Yrjö Häme, yh2475@columbia.edu, Columbia University, USA.

At what point does the human diagnostician's eye no longer remain the "gold standard"?

Huddled up around a single monitor in a dimly-lit reading room, four clinicians had their eyes fixated on computed tomography images of the chest. The volumes depicted patients suffering from pulmonary emphysema and the doctors were used as expert observers in a study for establishing baseline image biomarkers for different subtypes of the disease.

At the display of each new image, an argument ensued. The doctors could not agree on the subtype or degree of the disease, and even the desirable viewing conditions were a point of contention. Of course, this did not come as a surprise to anyone involved with the experiment. With typical inter-observer agreement values being around 0.5 [1], it would have been suspicious had the doctors agreed.

Computer vision methods in the medical domain are often evaluated against an expert reference, with success determined by the accuracy at which the method can repeat the expert's diagnosis, delineation or other assessment. In supervised methods, the reference data is even used to train the system. However, more often than a layman might think, a clinician's assessment is inaccurate and highly subjective. This inaccuracy can be due to multiple reasons, such as lack of experience, difficulty in taking all the data into consideration (especially in multimodal or functional data) or simply because even doctors get tired. In some cases, the underlying condition may not currently be known well enough to make a reliable assessment (e.g. [2]).

As exemplified by the story, the "gold standard" assessment provided by a clinician can be very different from the actual *ground truth*. Naturally, the ground truth is usually not available when dealing with patient data. From a computer vision researcher's standpoint, important questions arise: How can the accuracy of a method be evaluated if the reference data cannot be trusted? Should any of the reference data be used for training supervised methods, considering that inaccurate labels will bias the learning? And perhaps most importantly, how can a method be shown to perform more accurately than a diagnostician?

Showing that a computer vision system provides better repeatability or reduces manual labor compared to manual observation is usually fairly simple. On the other hand, improvements in accuracy can be very difficult to prove, even though it seems obvious that in many applications a computational model may be more accurate than the assessment of an average diagnostician. This is especially true for complicated and large datasets, that often include 3D images from multiple modalities, some of which may be functional.

To show increased accuracy, artificially created phantom data (e.g. [3]) is sometimes used to study accuracy when ground truth is available. Unfortunately, realistic phantom data is unavailable in many applications, or is too limited to represent the variability that is present in a real population.

Together with the proliferation of digital imaging, the constantly increasing capacity in computational resources and data storage have recently provided researchers with rich datasets that have previously been unavailable. This may be at least a partial solution to the problem. Longitudinal imaging datasets, where images of the same subject are collected at different time points, introduce great potential to computer vision research. In addition to imaging data, also the outcome of the condition or the current status of the subject is often available.

There are multiple related approaches for using longitudinal data to prove that a computer vision system should be considered the "gold standard" instead of a clinician's assessment. The following tests are some of the alternatives that are readily usable for credible comparison: First, if a computer vision system can show a correct diagnosis at an earlier stage than diagnosticians, the computer vision system should be considered a more accurate approach. Second, if a predictive model is used to estimate a subject's condition at a later stage, the assessment providing measurements that result in more stable and reliable prediction should be considered superior. Third, related to the second alternative, the accuracy of the prognosis provided by each approach can be used for comparison.

These evaluation alternatives might not be as simple to implement as the standard sensitivity and specificity measures that are commonly used. However, they will most likely be used as standard evaluations for many applications within the next decade.

Even if the computer vision system can be shown to be superior by average accuracy, acceptance as a clinical "gold standard' is naturally another issue. It does not seem enough that a computer vision system performs better on average than an average clinician. Wide acceptance will most likely require showing performance that is without a doubt superior to assessments made by clinicians. Before that, the computer vision models will remain in a supporting role, helping diagnosticians by giving a second opinion, which is especially valuable for inexperienced doctors.

Medical applications are inherently multiscale problems, where not only the imaging data is considered for the final decision. Therefore to be a true gold standard, all of the information that a diagnostician has should also be incorporated to the image-based analysis.

References

[1] Bankier, A.A., De Maertelaer, V., Keyzer, C., Gevenois, P.A.: Pulmonary Emphysema: Subjective Visual Grading versus Objective Quantification with Macroscopic Morphometry and Thin-Section CT Densitometry. Radiology, vol. 211, 1999.

[2] Curkendall, S.M., deLuise, C., Jones, J.K., Lanes, S., Stang, M.R., Goehring, E., She, D.: Cardiovascular disease in patients with chronic obstructive pulmonary disease, Saskatchewan Canada: cardiovascular disease in COPD patients. Annals of epidemiology, vol. 16, 2006.

[3] Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., Evans, A.C.: Design and construction of a realistic digital brain phantom. IEEE Transactions on Medical Imaging, vol. 17, 1998.