35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Towards A Multi-Site International Public Dataset For The Validation Of Retinal Image Analysis Software*

Emanuele Trucco and Alfredo Ruggeri, *Senior Member, IEEE*

Abstract— **This paper discusses concisely the main issues and challenges posed by the validation of retinal image analysis algorithms. It is designed to set the discussion for the IEEE EBMC 2013 invited session "From laboratory to clinic: the validation of retinal image processing tools". The session carries forward an international initiative started at EMBC 2011, Boston, which resulted in the first large-consensus paper (14 international sites) on the validation of retinal image processing software, appearing in IOVS. This paper is meant as a focus for the session discussion, but the ubiquity and importance of validation makes its contents, arguably, of interest for the wider medical image processing community.**

I. Introduction and Background

At EMBC 2011, Boston, a group of researchers started a broad discussion on the validation of retinal image analysis (RIA) algorithms [1,2,6]. This discussion generated a joint paper capturing the state of the art and the consensus of 14 international sites on guidelines for the validation of RIA software [3]. To our best knowledge, this was the first joint effort from a substantial number of international groups to agree on procedures for RIA software validation.

Although validation is a complex and articulated topic, it is arguable that the single most important issue to enable international research is *the creation of public data repositories*. This, in turn, requires careful design and substantial resources [3,7,8], and solid collaborations with dedicated clinicians.

This paper is designed for the invited session "From laboratory to clinic: the validation of retinal image processing tools" at IEEE EBMC 2013, which aims to take forward in a public forum the collaborative action started in 2011. We suggest that the logical next step of the initiative started in 2011 is the creation of an international, multi-site data repository following the guidelines in [3].

To prepare the ground, this paper summarizes the key issues and challenges related to the validation of RIA tools, primarily the creation of an international pilot data set built according to the recommendations laid out in [3]. Its contents and suggested methodology are arguably of interest well beyond the RIA community. In addition, the paper reports briefly two validation-related experiences in the authors' groups. The ultimate vision is to help the research community to devise increasingly reliable RIA tools, supporting the shift of RIA tools from the laboratory to the clinic.

*Trucco's research is partially supported by Leverhulme Trust grant RPG-419 and SINAPSE/OPTOS project CARMEN (retinal biomarkers for cardiovascular conditions).

E. Trucco is with the VAMPIRE-CVIP group, School of Computing, University of Dundee, Dundee DD1 4HN, UK (phone: +44-1382-385504; fax: +44-1382-385509; email: manueltrucco@computing.dundee.ac.uk).

A. Ruggeri is with the Department of Information Engineering, University of Padova, 3531 Padova, Italy (email: alfredo.ruggeri@unipd.it).

II. Related Work

A. Validation in medical image processing

For our purposes, *validation* can be defined as *the process of showing that an algorithm performs correctly by comparing its output with a reference standard* [3]. In other words, validation is the experimental process by which a medical image processing (MIP) system is shown to achieve its purpose (e.g., locating an organ in MRI data, estimating the width of arteries in a specific region of a fundus colour image) to a certain quantitative extent, established, e.g., by ROC analysis or confidence levels of statistical tests.

The three basic needs of validation are perfectly captured by Jannin [7]: *standardization of validation methodology (protocols), design of public data sets*, *standardization of validation metrics*. This paper focuses on the second need. As any MIP system needs validation, and as there may be significant differences in validation requirements across clinical domains and applications (e.g., therapy, biomarkers, intervention, screening), the literature is largely fragmented. Topic-specific reports include the 2002 and 2006 TMI special issues [7,8]; forums include the working group on MIP within the European Federation for Medical Informatics [9], the Validation in Medical Image Processing initiative [10], and the Quantitative Imaging Network [11].

B. Validation in Retinal Image Processing

The validation of RIA software follows the general definition above but introduces domain-specific issues.

We summarize in a concise list the main RIA-specific challenges (see [3] for a comprehensive discussion).

1. The notorious *variability of expert judgment* [12,13] is countered by having multiple experts annotate the same data set. As it is impossible to claim an accuracy higher than that of the reference standard used, variations among experts must be characterized quantitatively. However there is no *ultimate* consensus on how to reconcile multiple reference values (e.g., averaging, discussion and consensus, inter-rater reliability metrics such as AC1 or Kappa, histograms, distributions).
2. *Annotation protocols*. Annotating specific image elements, like circling on a computer screen, is a task that

doctors do not not normally perform. To circumvent the ensuing problems (see Section 4.1), one ought to align validation and clinical tasks as much as possible. Quellec et al. [12] designed GUIs to automatically detect elements that catch the attention of clinicians in their daily clinical practice. This avoids requesting clinicians to annotate explicitly anatomical structures, a task they have not been trained for. Procedures used to take photographs represent another source of variability due to protocols.

3. *Generating annotations directly comparable to software output*. This is an obvious requirement, but has the drawback that some annotations tasks are not part of normal clinical practice (e.g., estimating accurately the width of blood vessels at many locations in a fundus images). Multiple-observer annotations may be confused by the fact that clinicians are not used to the task, or do not see its relevance. For these reasons some authors have begun to explore alternative paradigms, for example weak learning methods [23] (moving from algorithm-oriented annotations to the use of clinical notes directly) and STAPLE [24] (addressing the *simultaneous* reliability estimation of algorithm *and* reference standard from annotations by multiple experts).

3. *Outcome point*. It is not always clear where to set the "outcome" for validation. In screening programs, a "refer-no refer" decision seems the obvious choice; other cases are not so clear.

4. *Physiological short-term changes.* Taking photographs at random instants in the pulse cycle may result in unrecognized variations in the measurements of retinal vessel diameters, but no firm conclusions seem possible from the few studies reported so far [13,14].

5. *Different imaging instruments.* The level of customization for each RIA modality is high and algorithms suitable to one type of image may not be directly usable for a different type. Even within the same class of machines, instrument variation can have a large effect on algorithm performance, e.g., variations of resolution, FOV, color calibration model.

6. *Data/image quality.* Image quality depends on instrument characteristics, acquisition procedure, and target conditions. Quality definitions applied by experts are elusive to quantitative rules. In general, images suitable for clinical analysis may not produce good results with RIA systems.

7. *Data sets.* Dafferent data sets may lead to somewhat inconsistent performance assessments, as preparation protocols may differ. The design of data sets for RIA validation is a crucial issue (Section 1); the next section concentrates on it, and this will be tabled as an initiative proposal at the invited session chaired by the authors. Among the most popular, current public data sets with annotations for RIA, we mention STARE [15], DRIVE [16] (vasculature detection), REVIEW [17] (vessel width estimation), MESSIDOR [18], and the diabetic retinopathy online challenge [19] (DR-specific lesion detection).

III. Recommendations for Test Data Repositories

The creation of substantial, structured, public data sets built and certified by large groups of RIA researchers and clinicians would be a substantial push towards the development of RIA software tools closer to translation.

We summarize the criteria agreed in [3] for such data sets.

(a) Created collaboratively by consortia of international groups to achieve size and multiple annotators, to reduce opinion bias, and support international visibility and credibility.

(b) Easily accessible, ideally by suitably structured websites.

(c) Regularly maintained, to manage distribution, additions, and potential obsolescence of data and annotations.

(d) Large size; tentatively, the minimum order of magnitude should be the thousands of images.

(e) Include standardized, patient friendly imaging protocols allowing large populations to be imaged effectively.

(f) Include metadata, i.e., non-image data characterizing imaging instruments, patients and disease.

(g) Include automated tools for running software on the data, as done, e.g., by the Middlebury stereo site [25], in which executable code is loaded and run on the site, and performance assessed in terms of pre-defined measures which are displayed in tabular form.

(h) Organized by outcome, which depends on the task at hand. An image set could be used for multiple outcomes by providing multiple types of annotations.

(i) Include image annotations, providing the standard reference for comparison for the outcome stated, by as many clinicians as possible (ideally from different sites to eliminate possible opinion bias); each expert should ideally annotate the data set multiple times to estimate intra-observer variability.

IV. Experiences in RIA Validation

To illustrate the diversity of the challenges of RIA validation, we report briefly some experiences from the authors' groups.

A. RIA Validation and Software Engineering

VAMPIRE is an international collaboration of 10 clinical and image processing centers [4,5] developing a software suite for the efficient quantification of morphological features of the retinal vasculature in large sets of fundus camera images. The measurements are intended mostly for biomarker discovery, see for instance [9, 10]. We share here two points emerging from the VAMPIRE experience so far, exposing, in our view, the intimate connection of validation and software engineering.

(a) The design of annotation tools. Like other groups, we have created specific, interactive software tools to enable clinicians to generate ground truth annotations. These are used to validate segmentation and location algorithms (e.g., optic disc, fovea center, vessel junctions and width). Even for deceptively straightforward tasks like vessel width estimation, however, we found ourselves having to add or modify repeatedly the GUI and the data saved, to address asynchronous suggestions from clinical annotators. The key fact seems to be that the interpretation of an annotation task (e.g., vessel width estimation at specific locations) is

different for the image processing expert ("clicking vessel contour points") and for the clinician, who tends to relate the annotation to his clinical background and tasks.
Requirement collection is therefore crucial in the development of annotation tools. We learnt to allow ample time for it, to involve multiple clinicians, and to discuss a variety of tasks for which the annotations can be used.

(b) Data repository and data engineering. The unfolding of a continuing RIA research program brings about novel directions of investigation. These, in turn, generate new data. In our biomarker-related work, we started to operate in a *software-centered mode*, in which algorithm development was the key operation, and data (images) important as satellite entities (Figure 1). It is now apparent to us, however, that *validated results can become data themselves*, e.g., taken as *data* for further clinical studies, or simply re-used in extended-functionality tool. This suggests strongly a *data-centric approach*, in which a data repository is "surrounded" by a cluster of software modules which augment the repository with new data. The data repository becomes therefore a *dynamic entity*, involved in a bidirectional interaction with the software cluster. This is, in essence, the model behind DICOM (medical.nema.org) for radiological images and, *mutatis mutandis*, and the Open Microscopy Environment for life sciences (OMERO, www.openmicroscopy.org). VAMPIRE is currently being aligned with this vision.

B. RIA Validation Criteria for Tortuosity Estimation

An important methodological issue for validation arises when a property of clinical interest is not associated with numerical values. This case was dealt with at LBI-UoP, addressing *tortuosity*. This geometric feature of vessels plays an important role in the diagnosis and grading of several retinal and vascular diseases, but there is no ultimate consensus on its numerical quantification (unlike, e.g., vessel width or disc size).
The LBI-UoP group recently developed algorithms to automatically measure tortuosity of retinal vessels in adults images [20], in infants images acquired with a wide-field fundus camera [21], and of corneal nerves in images acquired with confocal microscopy [22]. To derive an annotated dataset to be used as ground-truth for the validation of our tortuosity algorithms, we decided to acquire, for each of the three applications, a sizable set of images and to collect and annotate the experts assessment in two different ways.
For retinal images [20,21], we asked the experts to order all the images in the dataset by increasing perceived tortuosity, using a side-by-side comparison between pairs of images. The validation was then performed by measuring the Spearman rank correlation coefficient between expert and automated image ordering.
For corneal nerve images [22], we asked the experts to classify all the images in the dataset as having “low”, “mid” or “high” tortuosity. We then evaluated the tortuosity algorithm by measuring the classification error and Krippendorff concordance coefficient it obtained with respect to the ground-truth expert classification.

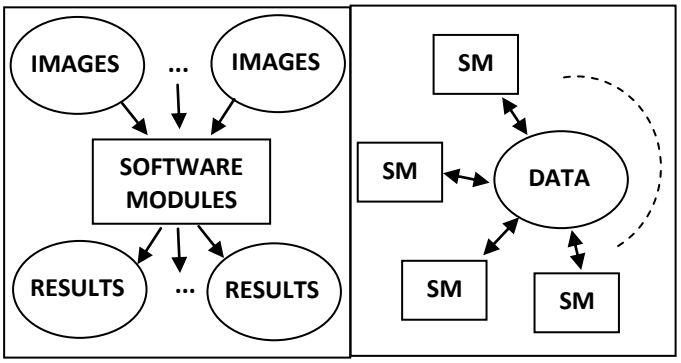

Figure 1. Left: software-centred model: data and results are separate and static. Right: data-centred model: data is a dynamic entity continuously augmented by results, which become themselves new data.

In these instances the validation procedure had to cope with the limitations of human expert assessment and thus had to resort to unusual or categorical comparisons to evaluate the performance of RIA software.

V. DISCUSSION

This paper has deliberately no conclusions as it is meant to set the ground for the discussion at the IEEE EMBC 2013 invited session "From laboratory to the clinic: the validation of retinal image processing tools". Along with presentations by leading RIA groups, we shall propose the creation of an international pilot data set built according to the recommendations laid out in [3]. Its contents and suggested methodology are arguably of interest well beyond the RIA community. Ultimately, the intent is to help the research community to devise increasingly reliable RIA tools, supporting the shift of RIA software from the laboratory to the clinic.

ACKNOWLEDGMENT

This paper stems from the work of a large number of people. They include all authors and acknowledged contributors of [3]; the VAMPIRE groups in Dundee (L Ballerini, K Zutis, E Pellegrini, Dr P J Wilson, Dr A Doney), Edinburgh (T MacGillivray, D Relan, G Robertson, Dr B Dhillon), Palermo (C Lupascu, M Tegolo) and Verona (A Giachetti), and collaborators in Singapore (Jimmy Liu and team, ASTAR; Dr Augustinus Laude, Tan Tock Seng hospital), USA (Dr J P Hubschman and team, UCLA). We thank OPTOS plc (www.optos.com) for their continued commitment and support, and C Robertson (Epipole plc) for software advice.

REFERENCES

[1] M.D. Abràmoff, M.K. Garvin, and M. Sonka, *Retinal imaging and image analysis*, IEEE Reviews in Biomedical Engineering, 3, 169–208, 2010.

[2] N Patton, T.M. Aslam, T. MacGillivray, I.J. Deary, B. Dhillon, R.H. Eikelboom, K. Yogesan, et al., *Retinal image analysis: concepts, applications and potential,* Progress in Retinal and Eye Research, 25(1), 99–127, 2006.

[3] E. Trucco, A. Ruggeri, T. Karnowski, L. Giancardo, E. Chaum, J.P. Hubschman, B. al-Diri, C. Cheung, D. Wong, M. Abràmoff, G. Lim, D. Kumar, P. Burlina, N. Bressler, H. Jelinek, F. Meriaudeau, T. MacGillivray, B. Dhillon, *Validating retinal fundus image analysis algorithms: issues and a proposal*. Investigative Ophthalmology and Visual Science, in press, 2013.

[4] E. Trucco, L. Ballerini, D. Relan, A. Giachetti, T. MacGillivray, K. Zutis, C. Lupascu, D. Tegolo, E. Pellegrini, G. Robertson, P. J. Wilson, A. Doney, B. Dhillon, *Novel VAMPIRE algorithms for quantitative analysis of the retinal vasculature*. Proc. IEEE Biosignals and Biosystems, Rio de Janeiro, Feb 2013.

[5] A. Perez-Rovira, T. MacGillivray, E. Trucco, K.S. Chin, K. Zutis, C. Lupascu, D. Tegolo, A. Giachetti, P.J. Wilson, A. Doney, B. Dhillon, VAMPIRE: Vessel Assessment and Measurement Platform for Images of the REtina. Proc 33rd IEEE EMBC, Boston (USA), 2011.

[6] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman, Blood vessel segmentation methodologies in retinal images: a survey, Comput. Methods & Programs in Biomed., 108(1), 407-33, 2012.

[7] P. Jannin: *Validation of medical image processing in image-guided therapy*. Transactions on Medical Imaging, 21(12), 1445-1449, 2002.

[8] P. Jannin, E. Krupinski, E. Warfield, *Validation in medical image processing*. Guest editorial, special issue on validation in MIP, IEEE Transactions on Medical Imaging, 25(11), 2006.

[9] F. N. Doubal, T. J. MacGillivray, P. E. Hokke, B. Dhillon, M. S. Dennis, and J. M. Wardlaw, *Differences in retinal vessels support a distinct vasculopathy causing lacunar stroke*, Neurology, vol. 72, pp. 1773–1778, 2009.

[10] N. Patton, T. Aslam, T. J. MacGillivray, A. Pattie, I. J. Deary, and B. Dhillon, *Retinal vascular image analysis as a potential screening tool for cerebrovascular disease*, Journal of Anatomy, vol. 206, pp. 318–348, 2005.

[11] M.D. Abramoff, J.M. Reinhardt, S.R. Russell, J.C. Folk, V.B. Mahajan, M. Niemeijer, and G. Quellec. *Automated early detection of diabetic retinopathy*. Ophthalmology, 117(6):1147-1154, 2010.

[12] G Quellec, M Lamard et al., *Automated assessment of diabetic retinopathy severity using content-based image retrieval in multimodal fundus photographs*. Invest Ophthalmol Vis Sci. 2011, 52(11):8342-8.

[13] F. Moret, C. Poloschek, W. Lagrèze, and M. Bach, *Visualization of fundus vessel pulsation using Principal Component Analysis*, Investigative Ophthalmology & Visual Science, vol. 52, pp. 5457-5464, July 2011

[14] D K Kumar, H Hao, B Aliahmad, T Wong, R Kawasaki, *Does Retinal Vascular Geometry vary with Cardiac Cycle?* Investigative Ophthalmology and Visual Science 2012, 53(9):5799-805.

[15] www.ces.clemson.edu/~ahoover/stare/

[16] www.isi.uu.nl/Research/Databases/DRIVE/

[17] http://reviewdb.lincoln.ac.uk/

[18] http://messidor.crihan.fr/download-en.phz

[19] M. Niemeijer, B. Ginneken, M.J. Cree et al., *Retinopathy Online Challenge: Automatic Detection of Microaneurysms in Digital Color Fundus Photographs*, IEEE Transactions on Medical Imaging, 29(1): 185-195, 2010.

[20] E. Grisan, M. Foracchia, A. Ruggeri. *A novel method for the automatic grading of retinal vessel tortuosity*. IEEE Trans Med Imag 27(3), 310-9, Mar 2008.

[21] E. Poletti, E. Grisan, A. Ruggeri. *Image-level Tortuosity Estimation in Wide-field Retinal Images from Infants with Retinopathy of Prematurity*, Proc. 34th Annual International Conference of IEEE-EMBS, pp. 4958-61, IEEE, New York, 2012.

[22] F. Scarpa, X. Zheng, Y. Ohashi, A. Ruggeri. *Automatic Evaluation of Corneal Nerve Tortuosity in Images from In-vivo Confocal Microscopy*, Invest Ophthalmol Vis Sci , 52(9), 6404-8, 2011.

[23] G. Quellec, M. Lamard, G. Cazuguel, M.D. Abràmoff, B. Cochener, C. Roux, *Weakly Supervised Classification of Medical Images*, Proc. IEEE ISBI, 2012.

[24] O. Commonwick,, Warfield, D. Simon, *Estimation of Inferential Uncertainty in Assessing Expert Segmentation Performance From STAPLE*, IEEE Trans on Medical Imaging 29(3), 771-780, 2010.

[25] http://vision.middlebury.edu/stereo/data