Estimating Maximal Measurable Performance for Automated Decision Systems from the Characteristics of the Reference Standard. Application to Diabetic Retinopathy Screening.

Gwénolé Quellec and Michael D. Abràmoff

Abstract— **We investigate the maximal performance that can be measured for automated binary decision systems in terms of area under the ROC curve (AUC), against a reference standard provided by human readers. The goal is to determine the required characteristics of the reference standard to assess and compare automated decision systems with a given degree of confidence, or, to determine what degree of confidence can be obtained given the characteristics of the reference standard. We modeled the expected value of the AUC that can be measured for a perfect decision system, given a reference standard provided either by a single human reader or by multiple human readers (consensus, majority vote). The proposed model was applied to diabetic retinopathy screening in a dataset of 874 eye fundus examinations graded by three readers. The expected value of the AUC for a perfect decision system was estimated at 0.956 against a single human reader, and 0.990 against a 'majority wins' vote of three human readers. The Iowa detection program has reached the maximal performance measurable by a single human reader (0.929, CI: [0.897-0.962]) and is close to the maximal performance measurable by a 'majority wins' vote (0.955, CI: [0.939-0.972]).**

I. Introduction

Automated decision systems are rapidly growing in importance, typically when the amount of information is so large that it cannot be processed exhaustively or efficiently by human readers. An important field is wide-scale, massive disease screening, where thousands or even millions of images must be evaluated, such as mammograms, colonoscopy images or diabetic retinopathy screening. In order to allow translation of such decision systems into clinical practice, their performance relative to human readers needs to be determined. In the image based computer aided detection field, human readings and annotations are often accepted at face value. But human readers, usually physicians specialized in that field, are not perfect. This is shown by the interobserver variability and the intraobserver variability. Given that the reference standard almost never represents the true state of the disease for all patients correctly, the central question is what can be measured given the characteristics of the reference standard, or, on the contrary, what characteristics does a reference standard require to allow reliable measurement of a desired performance? In this study, we focus on binary decisions (e.g. normal versus abnormal): the automated decision system assigns an abnormality probability to each patient, ranging from 0 (high confidence that it is normal) to 1 (high confidence that it is abnormal). The performance of automated binary decision systems is commonly measured using the Area Under the ROC (Receiver Operating Characteristic) Curve (AUC), which is considered the most accurate and comprehensive measure of performance for binary decision systems [1]. We propose to model the expected value of the maximal AUC that can be measured for an automated decision system, as a function of the characteristics of the reference standard. Specifically, we model the expected value of the AUC, measured in a dataset annotated by a single reader, for a hypothetical decision system that always assigns a higher abnormality probability to abnormal patients than to normal patients. Confidence in the AUC measured for an automated decision system is usually represented by confidence intervals. Several approaches have been presented in the literature to define a confidence interval on the AUC, using a parametric [2] or a non-parametric [3] model of the errors made by the automated decision system. But these systems do not model the variability among human readers.

This paper focuses on Diabetic Retinopathy (DR) screening. DR is the most common cause of blindness in the working age population of the United States and of the European Union [4]. In the last decade, many automated decision systems have been proposed to interpret digital photographs of the retina, in order to help early detection of DR [5]. Only a few detection systems have been assessed in large screening datasets [6], [7], [8]: in almost all studies, each exam in the dataset was read by a single human reader.

II. Modeling the AUC of a perfect automated binary decision system

Given the performance of human readers, we model the maximal measurable performance of automated decision systems, in a reference dataset annotated by a single human reader. Specifically, what we model is the expected value of the AUC we would measure for a perfect decision system in such a dataset.

A. Definitions

Let $\mathcal{D} = \{c_1, c_2, ..., c_n\}$ be a dataset consisting of cases $c_{i,i=1..n}$. Each case c_i in the dataset is associated with a

G. Quellec is with Inserm, UMR 1101, SFR ScInBioS, Brest F-29200, France gwenole.quellec@inserm.fr

M. D. Abràmoff is with the Department of Ophthalmology and Visual Sciences and the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA

M. D. Abràmoff is with the Department of Veterans Affairs, Iowa City VA Medical Center, Iowa City, IA 55242, USA

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

binary label l_i: l_i='false' if the case is thought to be normal, l_i='true' otherwise.

1) Label vector: vector $(l_i)_{i=1..n}$ consisting of the labels assigned by one reader (either an actual human reader or a committee human readers or a simulated human reader) to each case $c_{i,i=1..n}$ in the dataset.

2) Reference standard: label vector $(l_i)_{i=1..n}$ provided by a human reader (or by human readers) for each case $c_{i,i=1..n}$ in the dataset.

3) Reference gold standard: reference standard obtained by a committee of human readers, either by a consensus or by a 'majority wins' vote.

4) Observed prevalence: percentage of cases $c_{i,i=1..n}$ in the database whose label (that has been assigned by imperfect human readers) is l_i ='true'.

5) Probabilistic labels: probability $p_{i,i=1..n}$, assigned by an automatic decision system, that case $c_{i,i=1..n}$ is abnormal.

6) Perfect decision system: hypothetical decision system that provides a strict ordering of the cases $c_{i,i=1..n}$ from the most obviously normal (c_{most_normal}) to the most obviously abnormal ($c_{most\_abnormal}$): $p_{most\_normal} = 0$ and $p_{most\_abnormal} = 1$. The concept of obviously normal/abnormal cases is illustrated hereafter. Observations from *obviously normal* cases would clearly show to any expert that there is no pathology, while observations from less *obviously normal* cases would show patterns that appear pathological. Similarly, observations from *obviously abnormal* cases would clearly show many pathological patterns, whereas observations from less *obviously abnormal* cases would only show a subtle pathological pattern, or even no pathological pattern at all (e.g. if it is occluded). A perfect decision system always assigns a greater probability to abnormal cases than to normal cases. In other words, with a proper cutoff on the ordering provided by a perfect decision system, the true state of the disease can be obtained for each case in the dataset.

B. Modeling the errors made by a human reader

In order to compute the expected value of the performance measurable for a perfect decision system, against a human reader, we design a model of the errors made by human readers against that perfect decision system. Our goal is to design a model for an average human reader; as a consequence, the Probability Distribution Function (PDF) of false positives and of false negatives made by human readers are modeled by "smooth" curves. We assume that the probability of errors decreases as the cases become more obvious. By definition of a perfect decision system, the degree of obviousness of a case is directly related to the index of that case in the ordering provided by a perfect decision system. As a consequence, the PDF of false positives and of false negatives made by human readers are modeled by smooth monotonic functions of the index in that ordering. Cubic Bézier curves are used in this paper to generate smooth monotonic curves [9].

1) Error model (see Fig. 1): the cases are ranked by a perfect decision system along the x-axis, from the most obviously normal to the most obviously abnormal. The index of a case in the ordering is normalized to the real interval $[0;1]$, in order to make the model independent of the dataset cardinality. The probability that a human reader classifies a case x as normal if it is actually abnormal (false negative), or as abnormal if it is actually normal (false positive), is represented along the y-axis.

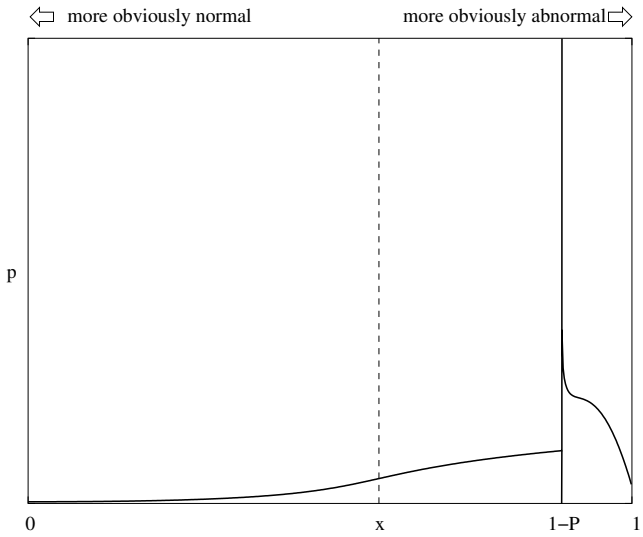

Fig. 1: Probability distribution function of the errors made by a human reader. The normalized index of all the normal cases, assigned by a perfect decision system, lie in the $[0; 1-P]$ interval, where P is the true prevalence. That of all the abnormal cases lie in the $]1-P; 1]$ interval.

2) Cubic Bézier curves: let B_0, B_1, B_2 and B_3 be four control points in $\mathbb{R}^2$. Given B_0, B_1, B_2 and B_3, a cubic Bézier curve $B(t), t \in [0;1]$, is a parametric curve running from B_0 to B_3 according to:

$$B(t) = (1-t)^3 B_0 + 3(1-t)^2 t B_1 + 3(1-t)t^2 B_2 + t^3 B_3 \quad (1)$$

The curve's derivative in B_0 is directed by B_0B_1 and its derivative in B_3 is directed by B_2B_3.

3) PDF of false positives and of false negatives made by human readers: let P be the prevalence and FPR_{HR} (resp. TPR_{HR}) be the average false positive rate (resp. true positive rate) of human readers. The false positive probability for a case $x \in [0; 1-P]$ is given by a Bézier curve P_{FP} meeting the constraint:

$$\int_0^{1-P} P_{FP}(x)dx = FPR_{HR} \quad (2)$$

The false negative probability for a case $x \in [1-P; 1]$ is given by another Bézier curve P_{FN} meeting the constraint:

$$\int_{1-P}^{1} P_{FN}(x)dx = 1 - TPR_{HR} \quad (3)$$

The observed prevalence, noted P_{HR}, is given by:

$$P_{HR} = \int_0^{1-P} P_{FP}(t)dt + \int_{1-P}^{1} (1 - P_{FN}(t))dt \quad (4)$$

To ensure that P_{FP} is a monotonically increasing function, its control points B_0^{FP}, B_1^{FP}, B_2^{FP} and B_3^{FP} have to meet the following constraints:

$$B_0^{FP} = (0, y_0), y_0 > 0 \quad (5)$$

$$B_3^{FP} = (1 - P, y_3), y_3 > y_0 \quad (6)$$

$$B_1^{FP} = (x_1, y_1), 0 \leq x_1 \leq 1 - P, y_0 \leq y_1 \leq y_3 \quad (7)$$

$$B_2^{FP} = (x_2, y_2), 0 \leq x_2 \leq 1 - P, y_0 \leq y_2 \leq y_3 \quad (8)$$

To ensure that P_{FN} is a monotonically decreasing function, its control points B_0^{FN}, B_1^{FN}, B_2^{FN} and B_3^{FN} have to meet the following constraints:

$$B_3^{FN} = (1, y_3), y_3 > 0 \quad (9)$$

$$B_0^{FN} = (1 - P, y_0), y_0 > y_3 \quad (10)$$

$$B_1^{FN} = (x_1, y_1), 1 - P \leq x_1 \leq 1, y_3 \leq y_1 \leq y_0 \quad (11)$$

$$B_2^{FN} = (x_2, y_2), 1 - P \leq x_2 \leq 1, y_3 \leq y_2 \leq y_0 \quad (12)$$

There are six parameters per curve (y_0, y_3, x_1, y_1, x_2, y_2) and 1 identity (see equations 2 and 3). As a consequence, there are five free parameters per curve.

C. Computing the maximal measurable performance from the human reader model

In order to compute the expected value of the AUC for a perfect decision system, we have to define, for any cutoff x (i.e. any case x), the True Positive Rate (TPR) and the False Positive Rate (FPR) of this decision system against a single human reader. Let $TPR_{DS}(x)$ be the expected value of the FPR, and $FPR_{DS}(x)$ be the expected value of the TPR, for any cutoff x. If $x \leq 1-P$, $TPR_{DS}(x)$ and $FPR_{DS}(x)$ are given by the following relations:

$$FPR_{DS}(x) = 1 - \frac{1}{1 - P_{HR}} \int_0^x (1 - P_{FP}(t))dt \quad (13)$$

$$TPR_{DS}(x) = \frac{1}{P_{HR}} \left(\int_{1-P}^1 (1 - P_{FN}(t))dt + \int_x^{1-P} P_{FP}(t)dt \right) \quad (14)$$

If $x > 1-P$, $TPR_{DS}(x)$ and $FPR_{DS}(x)$ are given by the following relations:

$$FPR_{DS}(x) = 1 - \frac{1}{1 - P_{HR}} \left(\int_0^{1-P} (1 - P_{FP}(t))dt + \int_{1-P}^x P_{FN}(t)dt \right) \quad (15)$$

$$TPR_{DS}(x) = \frac{1}{P_{HR}} \int_x^1 (1 - P_{FN}(t))dt \quad (16)$$

Finally, the expected value of the AUC for a perfect decision system is given by [1]:

$$E(AUC) = \int_0^1 TPR_{DS}(t) \frac{\partial FPR_{DS}}{\partial t}(t)dt \quad (17)$$

Equation 17 is approximated according to the trapezoidal rule [10].

D. Parameter estimation

Given a prevalence P and the average TPR/FPR of human readers (TPR_{HR}/FPR_{HR}), the proposed model has ten undetermined parameters: five parameters per Bézier curve. The parameters of the model are determined in a dataset where the level of agreement between human readers, measured by Cohen's κ [11], is known. The parameters are chosen such that, in this dataset, the difference between the observed average κ and the expected value of κ (§II-D.1) is less than the observed standard error of κ.

1) Expected value of κ between modeled human readers: Cohen's κ is defined as $\frac{P_a - P_c}{1 - P_c}$, where P_a is the probability of agreement and P_c is the probability of chance agreement between two human readers. Let A be the 2×2 agreement matrix, where $A_{0,0}$ denotes the probability that both human readers agree that a case is normal, $A_{1,1}$ denotes the probability that both agree that a case is abnormal, and $A_{1,0}$ (resp. $A_{0,1}$) denotes the probability that only the first (resp. the second) thinks a case is abnormal. P_a and P_c can be expressed as follows:

$$P_a = A_{0,0} + A_{1,1} \quad (18)$$

$$P_c = (A_{0,0} + A_{0,1})(A_{0,0} + A_{1,0}) + (A_{1,0} + A_{1,1})(A_{0,1} + A_{1,1}) \quad (19)$$

In our case, the expected value of A is given by the following equations:

$$A_{0,0} = \int_0^{1-P} (1 - P_{FP}(x))^2 dx + \int_{1-P}^1 P_{FN}(x)^2 dx \quad (20)$$

$$A_{0,1} = A_{1,0} = \int_0^{1-P} (1 - P_{FP}(x)) P_{FP}(x) dx + \int_{1-P}^1 (1 - P_{FN}(x)) P_{FN}(x) dx \quad (21)$$

$$A_{1,1} = \int_0^{1-P} P_{FP}(x)^2 dx + \int_{1-P}^1 (1 - P_{FN}(x))^2 dx \quad (22)$$

2) Acceptable human reader model and maximal measurable performance: several sets of Bézier curve parameters may lead to an acceptable human reader model, i.e. a human reader model explaining the desired κ. As a consequence, the maximal measurable performance is defined as the maximal AUC measured for the perfect decision method against an acceptable human reader model. Once the model parameters have been determined, the model is valid in any dataset regardless of the observed prevalence, provided that these are read by human readers with similar performance (i.e. similar average TPR and FPR), for example attained by similar levels of clinical experience. Only parameter P needs to be changed in equations 2 to 16.

E. Extension: performance of a perfect decision system against a 'majority wins' vote of three human readers

Now let us assume the same dataset has been read by three different experts and the reference gold-standard is defined as the 'majority wins' vote of the three label vectors. The performance of a perfect decision system against this

TABLE I: Performance of the Iowa Detection Program

reference standard	single reader				three readers	
	expert 1	expert 2	expert 3	average	consensus [12]	majority wins
measured performance	*0.940*	*0.926*	*0.921*	0.929	0.937	0.955
confidence interval [3]	*[0.918-0.962]*	*[0.902-0.950]*	*[0.897-0.945]*	[0.897-0.962]	[0.916-0.959]	[0.939-0.972]
maximal measurable performance	0.956				?	0.990

reference gold-standard can be modeled easily. Let p denote the probability that an expert assigns a positive label to some case c in the dataset. The probability that at least two readers assign a positive label to c is given by $p^3 + 3p^2(1-p) = p^2(3-2p)$. Note that simulating a consensus of three experts is much more complex and was not addressed in this paper.

III. Application to diabetic retinopathy screening

A. The Messidor2 dataset

Deidentified digital fundus color images of 1748 eyes in 874 people with diabetes were used. These images were acquired in three DR screening programs in France (Paris, Brest, St. Etienne) using a video 3CCD camera (Canon Europe BV) on a Topcon TRCNW6 nonmydriatic fundus camera (Topcon USA, Inc) with a 45° field of view centered on the fovea [12], [13].

B. Performance of human readers in the Messidor2 dataset

Each case (two fundus color images from a patient) was graded for retinopathy severity by 3 masked independent retinal specialists and regraded with adjudication until consensus. According to the reference gold-standard, defined as the adjudicated label vector, the observed prevalence was 21.7%. The average TPR (resp. FPR) before adjudication was 80.7% (resp. 2.3%) and the average κ was 0.822 [12].

C. Results

The parameters of the model described in section II were adjusted to match the properties of the reference gold-standard: P=21.7%, TPR=0.807, FPR=0.023 and κ=0.822. The expected value of the AUC, measured by one human reader for a perfect decision system, was estimated at 0.956 in the Messidor2 dataset. The expected value of the AUC, measured by a 'majority wins' vote of three experts, was estimated at 0.990. As a comparison, the performance of the Iowa Detection Program, against various reference standards, is reported in table I.

IV. Discussion and conclusions

The purpose of this study was to investigate the maximal performance that can be meaningfully measured for an automated decision system, given the characteristics of the reference standard provided by human readers. A model was proposed and applied to Diabetic Retinopathy (DR) screening in a large dataset. The results show that our automated DR detection system has reached the maximal measurable performance against a single human reader ($0.897 \leq 0.956 \leq 0.962$ — see table I). In fact, measuring performance improvements by automated DR detection systems has become impossible, if each examination is annotated by a single human reader. Second, the results show that using a committee of human readers may lead to an increase in the measureable performance level of automated DR detection: an AUC of 0.990 can be measured, as opposed to 0.956 against a single reader. In that case, measuring performance improvements is still possible ($0.990 > 0.972$ — see table I). We expect our approach to be helpful to both guide performance analysis of automated decision systems, as well as help guide development of reference (test) datasets for such systems. Specifically, in our domain, we expect it to motivate the development of a reference gold-standard in a larger DR detection dataset, which is a highly expensive, time-consuming endeavor [14].

References

[1] R. G. Lehr and A. Pong, "ROC curve," in *Encyclopedia of Biopharmaceutical Statistics*, S.-C. Chow, Ed., April 2003, pp. 884–891.

[2] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, pp. 839–843, September 1983.

[3] E. A. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, pp. 837–845, September 1988.

[4] D. C. Klonoff and D. M. Schwartz, "An economic analysis of interventions for diabetes," *Diabetes Care*, vol. 23, pp. 390–404, March 2000.

[5] M. Niemeijer, B. van Ginneken, M. J. Cree, A. Mizutani, and G. Quellec et al., "Retinopathy Online Challenge: automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans Med Imaging*, vol. 29, pp. 185–195, January 2010.

[6] M. D. Abràmoff, J. M. Reinhardt, S. R. Russell, J. C. Folk, V. B. Mahajan, M. Niemeijer, and G. Quellec, "Automated early detection of diabetic retinopathy," *Ophthalmology*, vol. 117, pp. 1147–54, June 2010.

[7] S. Philip, A. D. Fleming, K. A. Goatman, S. Fonseca, P. Mcnamee, G. S. Scotland, G. J. Prescott, P. F. Sharp, and J. A. Olson, "The efficacy of automated "disease/no disease" grading for diabetic retinopathy in a systematic screening programme," *Br J Ophthalmol*, vol. 91, pp. 1512–1517, November 2007.

[8] E. Decencière, G. Cazuguel, and X. Zhang et al., "Teleophta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, pp. 196–203, April 2013.

[9] G. Farin, *Curves and surfaces for computer-aided geometric design (4 ed.)*. Elsevier Science and Technology Books, 1997.

[10] X.-H. Zhou, N. A. Obuchowski, and D. K. McClish, *Statistical Methods in Diagnostic Medicine*. Wiley, April 2011.

[11] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, April 1960.

[12] M. D. Abràmoff, J. C. Folk, and D. P. Han et al., "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmol*, vol. 131, pp. 351–7, March 2013.

[13] G. Quellec, M. Lamard, G. Cazuguel, L. Bekri, W. Daccache, C. Roux, and B. Cochener, "Automated assessment of diabetic retinopathy severity using content-based image retrieval in multimodal fundus photographs," *Invest Ophthalmol Vis Sci*, vol. 52, pp. 8342–8, October 2011.

[14] E. Trucco, A. Ruggeri, and T. Karnowski et al., "Validating retinal fundus image analysis algorithms: issues and a proposal," *Invest Ophthalmol Vis Sci*, vol. 54, pp. 3546–59, May 2013.