The Bi-Gamma ROC Curve in a Straightforward Manner

Ehtesham Hussain*

Department of Statistics, University of Karachi, Karachi-75270, Pakistan

Abstract: In biomedical research, biomarkers (diagnostic tests) are used in distinguishing healthy and diseased populations. The effectiveness and accuracy of a biomarker generally assessed through the use of a Receiver Operating Characteristic (ROC) curve model, and its functional such as area under the curve (AUC). The parametric (smooth) ROC curves are obtained under the specific distributions assumptions. A resulting ROC curve model is the plot of sensitivity versus 1-specificity for all possible threshold values. Most popular and widely used ROC curve model is bi-normal ROC curve model under the assumptions of normality. When the biomarker results are continuous and positively skewed (non-normal). The gamma distribution is supposed to a flexible model for positively skewed measurements. In practice use of bi-gamma ROC curve model is hindered by the fact that ROC function cannot be written in closed-form.

The solution of the problem is to use transformed invariance property of ROC curve model. Which assumes that the test results of both diseased and healthy are normally distributed after some monotone transformation [1].

In this paper we propose a simple approximation solution for the problem mentioned in above lines using a normal approximation due to Wilson and Hilfertys [2]. Which is useful to approximate gamma distribution results with classical normal distribution based results.

Keywords: Sensitivity, Specificity, Receiver Operating Characteristic (ROC) curve, Normal distribution, Gamma distribution.

INTRODUCTION

In biomedical research, biomarkers (diagnostic tests) are used in distinguishing healthy and diseased populations. The effectiveness and accuracy of such biomarkers are generally determined through the use of a Receiver Operating Characteristic (ROC) curve model, and its summary measure, such as area under the curve (AUC).

The ROC curve model is a popular way graphically displaying the discriminatory accuracy of a biomarker for distinguishing between two populations. It has been used in many scientific areas such as radiology [3] psychiatry [4] epidemiology [5] manufacturing systems [6] and biomedical problems [7]. There are many excellent texts and review articles on the ROC curve. We refer the reader to papers [3, 8, 9] and books [9, 10].

In ROC analysis, a person is assessed as diseased (positive) if the tested marker value (level) is greater than a given threshold value, otherwise the subject is diagnosed as healthy (negative). The accuracy of any given threshold value can be measured by the probability of a true positive (sensitivity) and probability of true negative (specificity).

The ROC curve is plot of sensitivity (q(c)) versus 1-specificity, (1-p(c)) over all possible threshold values (c) of the marker. By definition, ROC curve is monotone increasing from 0 to 1 and invariant to any monotone transformation of the test results. And its often a curve with certain level of smoothness when test results from the diseased and non-diseased subjects follow continuous distributions.

Denote the distribution functions of the measurement of the healthy and diseased populations by F and G respectively. Then the sensitivity equals 1-G (c) and specificity equals F (c). The Receiver Operating Characteristic (ROC) is the plot of 1-F(c) and 1-G(c) as c varies. Bamber noted that area under this curve equal to $A = P(Y > X)$ [11]. A larger area indicates the marker discriminates well between the populations being compared.

The AUC can be interpreted as the probability that a randomly chosen diseased subject will have a marker value greater than that of randomly chosen healthy subject i.e. $P(Y > X)$. The Figure **1**, shows two typical ROC curves and the performance level that could be expected from guessing. The greater the area of unit square that lies below the ROC curve (by solid line curve) the greater the discriminate power of a biomarker (classifier).

The most popular and widely used parametric model is bi-normal ROC curve model, due to its ease, manipulation and estimation. The bi-normal ROC curve model assumes one normal distribution for diseased population and one normal distribution for healthy population, or after some transformation [12]. If a biomarker results are continuous and positively skewed (non-normal), the gamma distribution (which is positively skewed) provides a flexible model in different

*Address corresponding to this author at the Department of Statistics, University of Karachi, Karachi-75270, Pakistan; E-mail: ehussain@uok.edu.pk

ISSN: 1814-8085 / E-ISSN: 1927-5129/12 © 2012 Lifescience Global

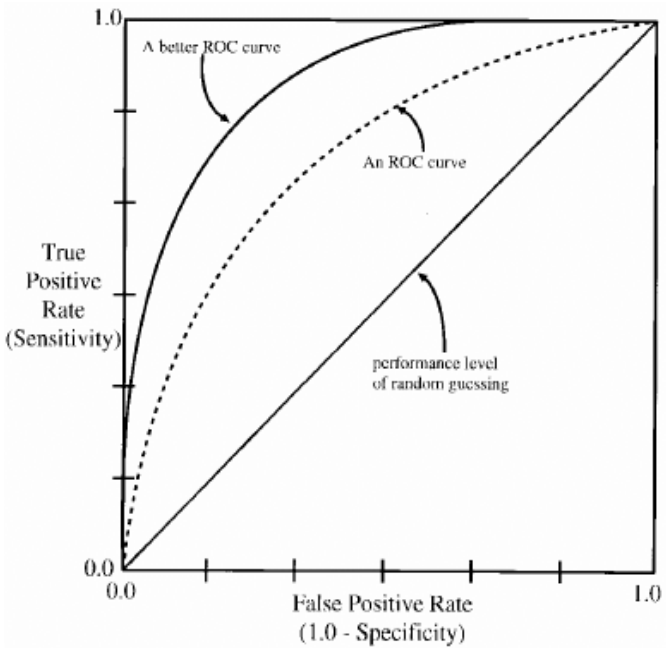

Figure 1: Two typical ROC curves and the performance level that could be expected from random guessing. The greater the area of the unit square that lies below the ROC curve, the greater the discriminate power of the biomarker (classifier).

situations. In practice use of bi-gamma ROC curve model is hindered by the fact that its ROC function cannot be written in closed form. To over come this problem we use the transformation invariant property of ROC curve. Which states that any ROC curve remains unchanged after a monotone transformation of the measurement of scale. In this paper we recommending the Wilson and Hilferty's transformation i.e. if X follows gamma distribution then $V = h(x) = X^{\frac{1}{3}}$ [2], follows normal distribution with mean $\mu_{1/3}$ and variance $\sigma^2_{1/3}$ i.e. $V \sim N\left(\mu_{1/3}, \sigma^2_{1/3}\right)$.

In this paper we first consider the bi-normal case and its relevant results. Next we study case of bi-gamma according proposed approach.

2.1. Bi-Normal ROC Curve Model

The bi-normal ROC curve plays a central role in ROC analysis. Which assumes one normal distribution for healthy population and normal distribution for diseased population. Assume that a biomarker's measures for healthy X, and diseased Y, are independent and normally distributed such that X~ $N(\mu_x, \sigma_x^2)$ and Y~N (μ_y, σ_y^2). Under these assumptions, sensitivity (q(c)) and specificity (p (c)) can be written as

$$q(c) = P(Y \geq c) = 1 - G(c) = 1 - \Phi\left(\frac{c - \mu_y}{\sigma_y}\right) \quad (1)$$

$$p(c) = P(X \leq c) = F(c) = \Phi\left(\frac{c - \mu_x}{\sigma_x}\right) \quad (2)$$

for a given cut-point c, where Φ denotes standard normal distribution function. The ROC curve is obtained by plotting q versus 1-p for all possible values of c i.e.

$$R(c) = \left\{1 - \Phi\left(\frac{c - \mu_x}{\sigma_x}\right), 1 - \Phi\left(\frac{c - \mu_y}{\sigma_y}\right)\right\} \quad (3)$$

Alternatively, given any common level of diagnostic specificity (1-p), then corresponding sensitivity of ROC curve is

$$q(p) = 1 - \Phi\left(\frac{\mu_y - \mu_x + \sigma_x \Phi^{-1}(1-p)}{\sigma_y}\right) \quad 0 < p < 1$$

$$q(p) = 1 - \Phi\left(\frac{\Phi^{-1}(1-p)}{\beta} - \frac{\alpha}{\beta}\right) \quad 0 < p < 1 \quad (4)$$

where $\alpha = \frac{(\mu_y - \mu_x)}{\sigma_x}$ and $\beta = \frac{\sigma_y}{\sigma_x}$.

Thus, the ROC curve is completely determined by the two parameters (α, β). Subsequently, inference can be made based on estimated parameters of (α, β). Furthermore the area under the ROC curve can be written as the probability that diseased individual has a higher biomarker value than diseased individual $A = P(Y > X)$ by definition it is

$$P(Y > X) = \int_{-\infty}^{\infty}\int_{x}^{\infty} f(x)g(y)dxdy$$

After some algebra one can find, the area under the ROC curve $A = P(Y > X)$ can be written as a closed-form solution for bi-normal distribution

$$A = \Phi\left(\frac{(\mu_y - \mu_x)}{\sqrt{(\sigma_y^2 + \sigma_x^2)}}\right)$$

$$A = \Phi\left(\frac{\alpha}{\sqrt{1 + \beta^2}}\right) \quad (5)$$

If the marker data is available on subjects from both the diseased and healthy populations, standard

estimates of the parameters $\mu_x, \mu_y, \sigma_x^2, and\ \sigma_y^2$ are available. Replacing the unknown parameters in the above formulae with their sample estimators $\overline{X}, \overline{Y}, S_x^2\ and\ S_y^2$ respectively the estimated parameters are

$$\hat{\alpha} = \frac{\overline{Y} - \overline{X}}{S_x}, \hat{\beta} = \frac{S_y}{S_x} \text{ and}$$

$$\hat{A} = \Phi\left(\frac{\hat{\alpha}}{\sqrt{1+\hat{\beta}^2}}\right) \quad (5.1)$$

2.2. Bi-Gamma ROC Curve Model

If a biomarker results are continuous and positively skewed (non-normal), the gamma distribution (which is positively skewed) provides a flexible model in many situations e.g. measurements of blood serum creatine kinase (CK) in Duchene muscular dystrophy (DMD) [13], pooled assessment of costly biomarkers, like interleukin-6 biomarker of inflammation for myocardial infection [14].

The bi-gamma ROC curve model involves two independent gamma distributions one for diseased population and one for healthy population. Now assume that a biomarker's measure for diseased, Y, and healthy, X, are not normally distributed but follow gamma distributions such that X~ $gamma(a_x, b_x)$ and Y~ $gamma(a_y, b_y)$, where

$$gamma(a,b) = \frac{e^{-x/b} x^{a-1}}{b^a \Gamma(a)} \quad (6)$$

where *x>0*, a>0 and b>0 are the shape and scale parameters respectively, and $\Gamma(a)$ is gamma function. For this distributional assumption

$$p = \frac{1}{\Gamma(a_x)(b_x)^{a_x}} \int_0^c x^{a_x-1} e^{-x/b_x} dx \quad (7)$$

$$q = \frac{1}{\Gamma(b_y)(b_y)^{a_y}} \int_c^\infty y^{a_y-1} e^{-y/b_y} dy \quad (8)$$

and the ROC curve can be obtained by graphing q versus 1-p for all possible c values.

Further, the area under the bi-gamma ROC curve can be shown [15] to be

$$A = \frac{1}{B(a_x, a_y)} \int_0^Q t^{a_y-1} (1-t)^{a_x-1} dt \quad (9)$$

where B(u,v) is the standard beta function and $Q = \frac{b_x}{b_x + b_y}$.

However, expression for (6), (7) for ROC curve and (8) for AUC are not as transparent because of the appearance of certain integrals. Seeing the above expressions (6), (7) and (8) it is obvious that they are complex and not in the closed form. In this situation invariant property of ROC curve is useful.

2.3. Invariant Property

A special feature of a ROC curve is that it is invariant to any monotone transformation of the measurement scale results, i.e. if V=h(x) and W=h(y) for some monotone (increasing) transformation h, then $p = F\left(h^{-1}(c')\right)$ and $q = \overline{G}\left(h^{-1}(c')\right)$. The transformation to normality usually is to be preferred, i.e. the theory developed for normal measurement scales also applies to transformed measurement scales.

2.4. Transformed Model

We now consider transformation [2]. Specifically, the Wilson and Hilfery approximation states if X follows a two-parameter gamma distribution [2], then the distribution of $X^{1/3}$ can be approximated by a normal distribution. For a gamma distribution with shape parameter 'a' and the scale parameter 'b', say gamma (a, b). The WH approximation now states that $X_{a,b}^{1/3}$ is approximately normal with mean and variance $X^{1/3} \sim N\left(\mu_{1/3}, \sigma_{1/3}^2\right)$ where

$$\mu_v = \mu_{1/3} = \frac{b^{1/3}\Gamma\left(a + 1/3\right)}{\Gamma(a)} \quad (10)$$

and

$$\sigma_v^2 = \sigma_{1/3}^2 = \frac{b^{2/3}\Gamma\left(a + 2/3\right)}{\Gamma(a)} - \mu_{1/3}^2 \quad (11)$$

The true ROC curve applying above transformation for both healthy and diseased results $V = X^{1/3} \sim N\left(\mu_v, \sigma_v^2\right)$ and $W = Y^{1/3} \sim N\left(\mu_w, \sigma_w^2\right)$ respectively.

Where $\mu_v, \mu_w, \sigma_v^2\ and\ \sigma_w^2$ are functional forms in (10) and (11). The true transformed ROC curve and area under the curve are:

$$R(c') = \Phi\left(1 - \Phi(c'), 1 - \Phi\left(\frac{c' - \alpha_1}{\beta_1}\right)\right) \quad (12)$$

Note that by invariant property area under the curve

A= $P(Y > X) = P(Y^{1/3} > X^{1/3}) = P(V > W)$.

$$A = \Phi\left(\frac{\alpha_1}{\sqrt{1+\beta_1^2}}\right) \quad (13)$$

where $\alpha_1 = \left(\frac{\mu_w - \mu_v}{\sigma_v}\right)$ and $\beta_1 = \frac{\sigma_w}{\sigma_v}$

To estimate parameters $\mu_{1/3}$ and $\sigma^2_{1/3}$ we ignore the functional forms in (10) and (11).

Thus if $X_1, X_2, \ldots, X_{n1}$ is a sample from a gamma(a,b) distribution, then we simply consider the transformed sample $V_1 = X_1^{1/3}, V_2 = X_2^{1/3}, \ldots, V_{n1} = X_{n1}^{1/3}$ as a sample from a normal distribution with an arbitrary mean μ_1 and arbitrary variance σ_1^2. Similarly if $Y_1, Y_2, \ldots, Y_{n2}$ is a sample from a gamma(c,d) distribution, then we simply consider the transformed sample $W_1 = Y_1^{1/3}, W_2 = Y_2^{1/3}, \ldots, W_{n2} = Y_{n2}^{1/3}$ as a sample from a normal distribution with an arbitrary mean μ_2 and arbitrary variance σ_2^2. The sample estimator for $\mu_1, \sigma_1^2, \mu_2, and\ \sigma_2^2$ are simply obtained by their natural sample estimators:

$$\overline{V} = \frac{1}{n_1}\sum_{i=1}^{n_1} V_i \quad (14.1)$$

$$S_v^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(V_i - \overline{V}\right)^2 \quad (14.2)$$

$$\overline{W} = \frac{1}{n_2}\sum_{i=1}^{n2} W_i \quad (14.3)$$

$$S_w^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(W_i - \overline{W}\right)^2 \quad (14.4)$$

respectively.

To estimate bi-gamma ROC curve model and its AUC from data one needs to estimate the means and variances separately from transformed data.

Using above simple sample estimators given in ((14.1), (14.2), (14.3) and (14.4)) then (12) and (13) become

$$\hat{\alpha}_1 = \frac{\overline{W} - \overline{V}}{S_v} \text{ and } \hat{\beta} = \frac{S_w}{S_v}$$

then estimate of a bi-gamma ROC curve and A are obtained in a straight forward manner.

2.5. Illustrative Example

In this section we consider hypothetical scenario for the gamma distribution. For simplicity, independence is assumed in over all individuals, both for healthy and diseased. The 20 Healthy biomarker levels (X) were generated from a gamma distribution with a=1.5 and b=1.0. The samples of 15 diseased levels (Y) were found from a gamma distribution with c=2.5 and d=1.25. Applying the WH approximation on X and Y to get transformed levels W and V respectively. Table **1**, shows, sample sizes, means and standard deviations needed to obtain bi-gamma ROC curve and AUC. The resulting bi-gamma ROC curve with an AUC=0.840 is presented in Figure **2**. This task was accomplished writing Mathematica code (as shown in appendix).

X(Healthy)= {1.9968,1.56302,2.51683,0.252915,0.215557,0.91478,0.222523,3.19765,1.35344,0.902746,0.811847,0.588331,0.919024,0.086737,2.18375,1.32928,2.63658,0.321508,0.804205,0.935241}

Y(Diseased) = {3.65447,2.50976,4.93018,1.08375,2.82772,4.13685,1.33389,1.91129,1.50153,1.68918,3.00398,0.679749,7.20512, 3.95123,6.3124}

V=($X^{1/3}$)= {1.25925,1.16053,1.36025,0.632399,0.59959,0.970746,0.60598,1.47325,1.10615,0.96647,0.932878,0.837929,0.972245,0.442658,1.29738,1.09953,1.38149,0.685063,0.929942,0.97793}

W=($Y^{1/3}$)= {1.54031,1.35897,1.70198,1.02717,1.4141,1.6053,1.1008,1.24101,1.1451,1.19095,1.44289,0.879258,1.93144,1.58092,1.84813}

Table 1: The Summary Results of Transformed Levels W, V

Variable	Sample size	Mean	Standard deviation
V	20	0.984582	0.289878
W	15	1.40055	0.30547

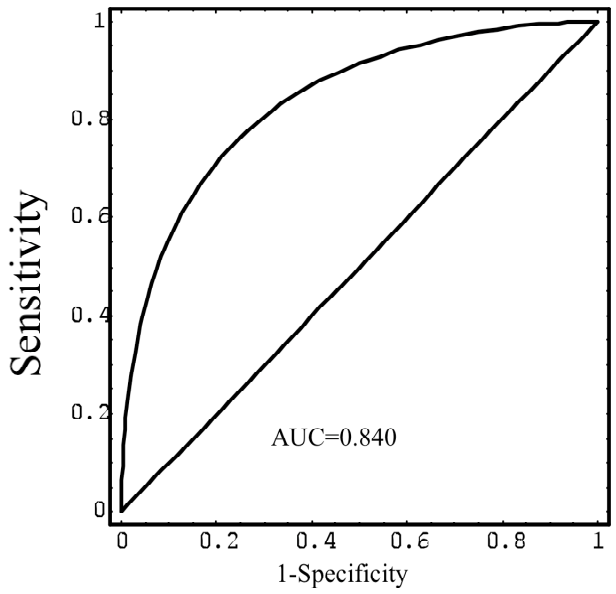

Figure 2: The Bi-gamma ROC Curve for parameters (a= 1.5,b=1.0.) and (c=2.5,d=1.25).

CONCLUDING REMARKS

The WH approximation is a simple normal approximation for a gamma distribution. In this paper we have exploited the approximation for ROC curve and its AUC problems for the gamma distribution. Because solutions are already available for the corresponding problems in normal case, the approximation has allowed us to adopt these solutions for the gamma distribution in straightforward manner.

APPENDIX:MATHEMATICA CODE

```
<<Statistics`ContinuousDistributions`
<<Statistics`DescriptiveStatistics`
                    (* Mathematica Code*)
(* First load following packages*)
(*<Statistics`ContinuousDistributions`

<<Statistics`DescriptiveStatistics` *)
(* This Programm generates gamma random
variables  and plots ROC curve  and
computes AUC *)
(* x1 healthy sample , y1 diseased
samples  *)
(* parmaters ax,bx, cy,dy, sample sizes
n1 and n2 repectivley *)

ClearAll
Clear[x1,y1, v,w,p,d,c,e,
ax,bx,cy,dy,n1,n2,r1,auc]
SeedRandom[10000];
(* Healty Sample, X1 *)
ax=1.5;bx=1.0 ;n1=20;
cy=2.5;dy=1.25;n2=15;
gx=GammaDistribution[ax,bx];
x1=RandomArray[gx,n1];
v=x1^(1/3);
mx=Mean[v]
sx=StandardDeviation[v]
gy=GammaDistribution[cy,dy];
SeedRandom[20000];
(* Dieased Sample, y1 *)
y1=RandomArray[gy,n2];
w=y1^(1/3);
my=Mean[w]
sy=StandardDeviation[w]
a=(my-mx)/sx
b=sy/sx
y[p_]:=N[(Quantile[NormalDistribution[0,
1],1-p]/b)-(a/b)];
d[e_]:=1-
CDF[NormalDistribution[0,1],y[e]];

p1=Plot[d[x],{x,0,1},AspectRatio®1/1,Fra
me®True,FrameStyle®{Thickness[0.01]},Axe
sLabel®{1-
sp,sen},PlotStyle®{Thickness[0.01]}];
p2=Plot[p,{p,0,1},AspectRatio®1/1,PlotSt
yle®{Thickness[0.01]}];
Show[{p1,p2}]
r1=a/Sqrt[1+b*b]
auc=N[CDF[NormalDistribution[0,1],r1]]

<<Statistics`ContinuousDistributions`
<<Statistics`DescriptiveStatistics`
                    (* Mathematica Code*)
```

REFERENCES

[1] Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. J Math Psych 1969; 6, 487-96. http://dx.doi.org/10.1016/0022-2496(69)90019-4

[2] Wilson EB, Hilferty MM. The Distribution of Chi-Squares, Proc Natl Acad Sci 1931; 17: 684-88. http://dx.doi.org/10.1073/pnas.17.12.684

[3] Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Investig Radiol 1989; 24: 234-45. http://dx.doi.org/10.1097/00004424-198903000-00012

[4] Somoza E, Mossman D, McFeeters L. The info-ROC technique: a method for comparing and optimizing inspection systems. In Review of Progress in Quantitative Nondestructive Evaluation, Thomson DO, Chimenti DE (eds). Plenum Press, New York 1990.

[5] Hsiao JK, Barko JJ, Potter WZ. Diagnosing diagnoses: receiver operating characteristic methods and psychiatry. Arch General Psychiatry 1989; 46: 664-67. http://dx.doi.org/10.1001/archpsyc.1989.01810070090014

[6] Aoki K, Misumi J, Kimura T, Zhao W, Xie T. Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogens I, II and of PG I/PG II ratios in a gastric cancer case-control study. J Epidemiol 1997; 7: 143-51. http://dx.doi.org/10.2188/jea.7.143

[7] Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Stat Sci Ser 2003.

[8] Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980s. Stat Med 1991; 10: 1887-95. http://dx.doi.org/10.1002/sim.4780101205

[9] Zhou H-X, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine, Wiley, New York 2002. http://dx.doi.org/10.1002/9780470317082

[10] Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press, Oxford 2003.

[11] Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 1975; 12 387-15. http://dx.doi.org/10.1016/0022-2496(75)90001-2

[12] Goddard MJ, Hinberg I. Receiver Operator Characteristic (ROC) Curves and Non-Normal Data: An Empirical Study. Stat Med 1990; 9: 325-37. http://dx.doi.org/10.1002/sim.4780090315

[13] Faraggi D, Reiser B. Estimation of the area under the ROC curve. Stat Med 2002; 21: 3093-6. http://dx.doi.org/10.1002/sim.1228

[14] Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. Stat Med 2003; 22: 2515-7. http://dx.doi.org/10.1002/sim.1418

[15] Constantine K, Karson M, Tse S. Estimation of P(Y<X) in the gamma case. Commun Stat Simul 1986; 15(2): 365-88. http://dx.doi.org/10.1080/03610918608812513

Received on 20-03-2012 Accepted on 22-04-2012 Published on 19-05-2012

http://dx.doi.org/10.6000/1927-5129.2012.08.02.09

2012 Ehtesham Hussain; Licensee Lifescience Global.
This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.