Research article

A machine-learning approach to the prediction of oxidative stress in chronic inflammatory disease

A. Magon de la Villehuchet[1], M. Brack[2,3], G. Dreyfus[1], Y. Oussar[1], D. Bonnefont-Rousselot[4,5], M.J. Chapman[2,3,6], A. Kontush[2,3,6]

[1]*École Supérieure de Physique et de Chimie Industrielles, ESPCI –Paristech, Laboratoire d'Électronique (CNRS UMR 7084), Paris, France*
[2]*UMR S551 'Dyslipoproteinemia and Atherosclerosis', Université Pierre et Marie Curie (Paris 6), Paris, France*
[3]*INSERM, UMR S551, Paris, France*
[4]*Department of Metabolic Biochemistry, La Pitié Hospital (AP-HP), Paris, France*
[5]*Department of Biochemistry, EA 3617, Faculty of Pharmacy, Paris Descartes University, Paris, France*
[6]*AP-HP, Groupe hospitalier Pitié–Salpétrière, Paris, France*

Oxidative stress is implicated in the development of a wide range of chronic human diseases, ranging from cardiovascular to neurodegenerative and inflammatory disorders. As oxidative stress results from a complex cascade of biochemical reactions, its quantitative prediction remains incomplete. Here, we describe a machine-learning approach to the prediction of levels of oxidative stress in human subjects. From a database of biochemical analyses of oxidative stress biomarkers in blood, plasma and urine, non-linear models have been designed, with a statistical methodology that includes variable selection, model training and model selection. Our data demonstrate that, despite a large inter- and intra-individual variability, levels of biomarkers of oxidative damage in biological fluids can be predicted quantitatively from measured concentrations of a limited number of exogenous and endogenous antioxidants.

Keywords: machine learning, neural networks, training, model selection, variable selection, oxidative stress, antioxidants, biological markers

Introduction

Epidemiological studies have revealed a close correlation between elevation in oxidative stress, attenuation of antioxidant defence systems and development of a wide range of chronic human pathologies, including atherosclerosis, neurodegenerative diseases, cancer, inflammatory diseases and diabetes.[1–4] Conversely, elevated levels of antioxidants are frequently associated with reduced prevalence of these diseases.[1–4] Finally, it is relevant that oxidative stress plays a key role in the aging process.[5] The clinical relevance of oxidative stress is further emphasised by the predominantly negative findings in recent large-scale studies of the relationship between antioxidant supplementation and incidence of cardiovascular disease and cancer.[6–12] Given the assumption that antioxidant supplementation may be beneficial primarily in subjects with elevated levels of oxidative stress, the inability of dietary antioxidants to reduce the incidence of

Correspondence to: Anatol Kontush, INSERM Unité 551, Pavillon Benjamin Delessert, Hôpital de la Pitié, 83 boulevard de l'Hôpital, 75651 Paris Cedex 13, France. Tel: +33 1 42177976; Fax: +33 1 45828198; E-mail kontush@chups.jussieu.fr
Received 18 July 2008, revised manuscript accepted 4 November 2008

DOI 10.1179/135100009X392449

cardiovascular disease and cancer in such individuals can be related to the lack of knowledge of baseline levels of oxidative stress in the respective cohorts.[13–15] The absence of such data may have resulted in antioxidant supplementation in subjects displaying normal levels of oxidative stress, and who would not be predicted to display further benefit. These findings demonstrate that knowledge of the oxidative status of a given individual might represent a key element in prevention of the progression of chronic human pathologies. At a cellular level, oxidative stress has its origin in a spectrum of oxidative systems, the most prominent of which are NADPH oxidase, myeloperoxidase, xanthine oxidase, lipoxygenase, nitric oxide synthase, cytochrome P450, the mitochondrial electron transport chain, ceruloplasmin and transferrin. Oxidative damage to biomolecules represents a major consequence of oxidative stress, resulting in the accumulation of oxidatively-modified proteins, lipids, carbohydrates and nucleic acids.[1,16,17] Such oxidatively-modified biomolecules typically display impaired functionality, thereby providing a mechanistic explanation for the pathological role of oxidative stress; levels of oxidized biomolecules are, therefore, considered as biomarkers of oxidative damage and represent highly relevant biomarkers of oxidative stress.[1,16,17] Oxidative stress can equally be assessed by a less direct approach involving determination of levels and/or activities of exogenous (*e.g.* vitamin C, vitamin E, carotenoids) or endogenous (*e.g.* glutathione, thiols, uric acid) antioxidants and/or antioxidative systems which protect functional biomolecules from oxidation.[1] Diverse forms of oxidative insult that occur *in vivo* result in distinct profiles of biomarkers of oxidative stress. The diversity of oxidative species implies that the choice of biomarkers that can be universally applied to characterise systemic oxidative stress in a living organism constitutes a major challenge. Comprehensive comparative studies addressing this issue have recently been initiated by the US NIEHS in an animal model of oxidative stress.[18,19] Biomarkers of oxidative stress are, however, characterised by strong cluster interdependence reflecting, for example, related protective pathways (*i.e.* protection of vitamin E by vitamin C, and protection of biomolecules by vitamin E); such interrelationships facilitate identification of robust biomarkers and suggest the possibility of mutual prediction of biomarker levels. In order to assess the profile of biomarkers of oxidative stress in a French population, the first Clinical Centre for Oxidative Stress in Paris was launched in 2002. More than 10 established biomarkers of oxidative stress were measured, including plasma, whole blood or urine levels of exogenous and endogenous antioxidants and biomarkers of oxidative damage. In the present investigation, advantage has been taken of a database of biochemical blood and urine analyses of individuals in a range of health conditions from healthy to strongly pathological. We have evaluated the feasibility of predicting levels of biomarkers of oxidative damage from measured levels of exogenous and endogenous antioxidants. We now describe the clinical database and protocols for measurement of biomarkers of oxidative stress, our approach to machine learning methods and finally the predictive power of our models.

Subjects and methods

Clinical database: contents and protocols

Subjects

The Clinical Centre for Oxidative Stress opened in Paris, France, in December 2002; by the end of 2005, profiles of biomarkers of oxidative stress were available in plasmas from 731 subjects (250 males, 481 females). Clinical and biological parameters were also measured in each subject. In 150 subjects, a second assessment of systemic oxidative stress followed within 4–6 months after the first visit. The majority of subjects presented clinically confirmed diagnoses as follows: cardiovascular disease (n = 136), psychiatric disease (depressive syndrome and anxious disorders; n = 98), neurodegenerative disease (Alzheimer's disease, Parkinson's disease and multiple sclerosis; n = 61) rheumatic disease (n = 34), infectious disease (HIV and hepatitis C; n = 28), cancer (n = 24) and endocrinological disease (thyroid dysfunction; n = 20). In 74 subjects, the simultaneous presence of multiple (two or more) pathologies was diagnosed; these subjects were considered as polypathic and excluded from statistical analyses. Subjects (n = 127) who contacted our Centre in the absence of any known symptoms and who were free of a clinical diagnosis were considered as healthy controls. The remainder of the subjects (n = 129) presented relatively rare pathological conditions (n < 20 for each specific disease) and were, therefore, excluded from statistical analyses.

Blood samples

Venous blood (20 ml) was taken from each subject after an overnight fast and immediately centrifuged at 3000 rpm for 10 min. EDTA and heparin plasma were isolated and immediately frozen at –80°C until analysis. Urine was collected on the same visit and used for biomarker analyses within 24 h.

Biomarkers of oxidative stress

The typical profile of biomarkers of oxidative stress included measurements of plasma, whole blood or urine levels of substances of exogenous origin (vitamin C, vitamin E, β-carotene, selenium, zinc, copper), of endogenous antioxidants (reduced and oxidized glutathione, thiols, uric acid) and of biomarkers of oxidative damage (oxLDL, antibodies against oxLDL, lipid hydroperoxides, 8-OHdG).

Determination of vitamin C

Vitamin C was spectrophotometrically measured in plasma stabilized with 10% metaphosphoric acid as the reduction of 2,6-dichlorophenolindophenol using a Perkin Elmer Lambda 40 spectrophotometer.[20]

Determination of vitamin E and β–carotene

Vitamin E and β–carotene were simultaneously determined by HPLC (Alliance Waters, USA) coupled to a diode array detector (PDA 2996, Waters, USA).[21] Plasma levels of vitamin E were normalized to total cholesterol, which was determined by a standard colorimetric kit containing cholesterol oxidase.

Determination of selenium, zinc and copper

Plasma levels of selenium, zinc and copper were measured using inductively coupled plasma-mass spectroscopy.[22]

Determination of reduced and oxidized glutathione

Reduced (GSH) and oxidized (GSSG) glutathione were measured in whole blood using a Bioxytech GSH/GSSG-412TM kit (OxisResearch, Portland, OR, USA). Initially developed by Tietze,[23] this method employs Ellman's reagent (5,5′-dithiobis-2-nitrobenzoic acid, DTNB) which reacts with GSH to form a product spectrophotometrically detectable at 412 nm. The thiol-scavenging reagent, 1-methyl-2-vinylpyridinium trifluoromethanesulfonate, was used to prevent oxidation of GSH to GSSG during sample processing. GSSG was calculated as the difference between total glutathione (determined after reduction of GSSG to GSH by glutathione reductase and NADPH) and GSH.

Determination of glutathione peroxidase (GPx) activity

GPx activity was measured in freshly isolated erythrocytes in the presence of reduced glutathione, NADPH, sodium azide, and glutathione reductase as a decrease in NADPH absorbance at 340 nm.

Determination of total thiols and uric acid

Total plasma sulfhydryl groups were determined spectrophotometrically at 412 nm after their reaction with DTNB.[24] Plasma urate was measured using a commercially available analytical test (Kodak Ektachem DT Slides, Eastman Kodak Company, Rochester, UK).

Determination of oxLDL

Levels of oxLDL were measured using a competitive enzyme-linked immunosorbent assay (ELISA) kit supplied by Immunodiagnostik (Germany; inter- and intra-assay coefficients of variation, 6.2% and 7.0%, respectively). Briefly, oxLDL from the sample competes with a fixed amount of oxLDL bound to the microtiter well for the binding of the specific biotin-labelled antibodies. After a washing step that removed unreacted sample components, the biotin-labelled antibody bound to the well was detected by HRP-conjugated streptavidin. After a second incubation and an additional washing step, the bound conjugate was detected by reaction with TMB. The reaction was stopped by adding acid to produce a colorimetric end-point that was detected spectrophotometrically.

Determination of antibodies against oxLDL

The titre of IgG antibodies against oxLDL was assessed with a commercial enzymatic immunoassay (Biomedica Gruppe, Vienna, Austria) using Cu^{2+}-oxidized LDL as an antigen (inter- and intra-assay coefficients of variation, 10.5%).

Determination of lipid peroxides

Lipid peroxides were assessed in plasma using an Oxystat spectrophotometric kit (Biomedica), which employs peroxide hydrolysis, by a peroxidase followed by reaction with TMB as a substrate, with detection at 450 nm.

Determination of 8-hydroxy-2′-deoxyguanosine

Competitive ELISA was used for the quantitative measurement of the oxidative DNA adduct 8-OHdG in fresh urine samples (Japan Institute for the Control of Aging, Japan). The concentration of 8OHdG was normalised to urine levels of creatinine (expressed as ng/mg creatinine).

Statistical machine learning

Introduction to statistical machine learning

Statistical machine learning encompasses a variety of mathematical and statistical techniques that aim at reproducing the learning abilities exhibited by humans or animals. In that context, a 'machine' should not be interpreted to represent a physical object, but rather a set of algorithms and procedures that are implemented on a computer. The mathematical foundations of statistical machine learning are described by Vapnik.[25] In the present article, we focus on the application of machine learning to the design of predictive models in the form of non-linear, parameterized functions. In other words, functions are postulated, that are expected to: (i) 'explain', in a statistical sense, the existing observations; and (ii) generalize to hitherto unknown situations (*i.e.* predict the outcome of future measurements). For present purposes, the postulated functions are neural networks, as described below. Training is the algorithmic procedure whereby the parameters of the postulated function are adjusted in order to fit the measurements present in a database termed a 'training set'. The training procedure used in the present study is described cursorily below. Finding the appropriate complexity for the model, given the available data, is a central task in statistical machine learning. The complexity of a model is defined specifically by its Vapnik–Cervonenkis dimension, which is generally an increasing function of the number of adjustable parameters of the model. If a model is insufficiently complex, it is unable to learn the training data: it has a high bias (*i.e.* the distance between the model and the 'true' regression function is large), but its variance (*i.e.* its sensitivity to the idiosyncrasies of the available training data) is low; conversely, if the model is too complex, it exhibits low bias (*i.e.* it adjusts very accurately to the training data or 'overfits' the data) but high variance (*i.e.* it depends strongly on the details of the training data, therefore on the noise present in it). Since the generalization error involves the sum of the bias (which decreases as complexity increases) and of the variance (which increases as complexity increases), there exists a complexity for which the generalization error is minimum. Therefore, models of increasing complexity are designed, their prediction error is assessed, and the model with optimal complexity is selected. The model selection method used in the present study is described later. Variable selection is also a key issue in statistical machine learning. The purpose of variable selection is to detect candidate variables that are not relevant for the task at hand; more specifically, the variables whose influence on the quantity to be modelled is smaller than the noise in the measurement of that quantity should be discarded. In most present-day models, the number of adjustable parameters is an increasing function of the number of variables in the model; therefore, the presence of irrelevant variables results in unnecessary model complexity, thereby increasing the probability of overfitting. The variable selection method used in the present study is described below. In traditional regression, a knowledge-based model of the process of interest is derived from first principles, and the parameters of the model have a physical (biological, chemical, *etc.*) significance, so that it is useful to estimate confidence intervals for the values of the parameters found by regression. In machine learning, there is no such thing as a 'true' model, so that, in most cases, the parameters have no specific physical (biological, chemical, *etc.*) meaning: the focus is on the prediction itself, so that it is essential to estimate confidence intervals for the predictions. The specific confidence interval used in the present study is defined later.

Postulated functions

In the present study, the postulated functions are neural networks (for introductory textbooks, see Bishop[26] and Dreyfus[27]). A neuron is a non-linear, bounded, parameterized function. The neural networks used in the present study are linear combinations of so-called 'hidden' neurons; such neural networks are termed 'feed-forward neural networks' or 'multilayer Perceptrons'. More specifically, in the present study, a neuron is an s-shaped ('sigmoid') function of a linear combination of its variables. The neuron computes the value of f defined as:

$$f = \tanh (\theta \times x) \qquad \text{Eq. 1}$$

where $\boldsymbol{\theta}$ is the vector of parameters (or 'synaptic weights') of the neuron, and $\mathbf{x}$ is the vector of variables, with an additional component, termed 'bias', which is equal to unity; therefore, if N is the number of variables, the size of $\mathbf{x}$ is N+1.

A 'feed-forward neural network' $g(\mathbf{x})$ is a linear combination of N_h 'hidden' neurons f_i (i = 1 to N_h) and of a constant equal to 1. We denote by $\boldsymbol{\Theta}_1$ the vector of parameters of the linear combination (of size N_h+1), by $\boldsymbol{\Theta}_2$ the (N+1, N_h) matrix whose columns are the parameters of the 'hidden' neurons, and by $\mathbf{f}$ the vector (of size N_h+1) of functions computed by the hidden neurons, with an additional component equal to 1. Then the 'neural' model is:

$$g(\mathbf{x}) = \boldsymbol{\Theta}_1 \cdot \mathbf{f}(\boldsymbol{\Theta}_2\mathbf{x}) \qquad \text{Eq. 2}$$

Feed-forward neural networks are frequently described pictorially as shown on Figure 1. Such neural networks are universal approximators: any continuous, differentiable function can be approximated, with arbitrary accuracy, by a neural network of the type described above, provided the number of its hidden neurons is large enough. Therefore, the complexity of a

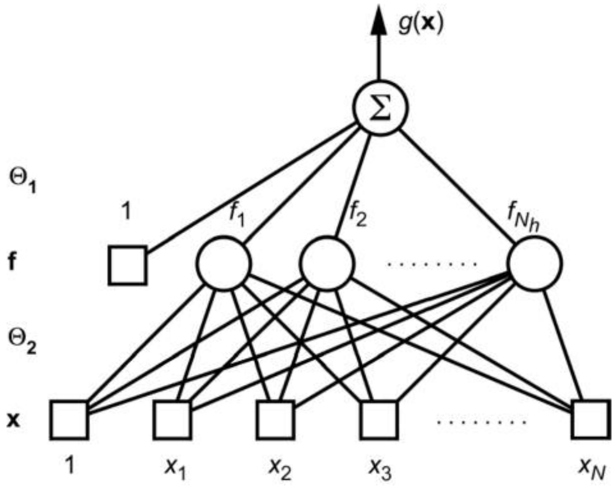

Figure 1 A feed-forward neural network with *N* variables and *N*$_h$ hidden neurons

neural network is determined by the number of hidden neurons N_h or, alternatively, by the number of parameters $(N+2)N_h+1$.

Neural networks are parsimonious: it is clear from Equation 2 that the model $g(\mathbf{x})$ is non-linear with respect to the parameters of matrix $\mathbf{\Theta}_2$, while a polynomial model, for instance, is linear with respect to all its parameters. In other words, a polynomial is a linear combination of monomials, whose shapes are fixed, while a neural network is a linear combination of functions whose shapes are adjusted during training; that additional flexibility decreases the requirement in terms of number of parameters. The number of parameters of a neural network varies linearly with the number of variables N, while the number of parameters of a polynomial increases as N^d where d is the degree of the polynomial; therefore, neural networks are less prone to overfitting than polynomial models and, more generally, than linear-in-their-parameters models, with the notable exception of Support Vector Machines, which have a built-in regularization mechanism. In order to demonstrate the feasibility of the predictions, using neural networks was deemed to be the most expedient way at the present stage of the investigation. A more detailed discussion of the pros and cons of the various machines lies beyond the scope of the present paper.

Training

Training was performed by minimizing the least squares cost function:

$$J(\mathbf{\Theta}_1,\mathbf{\Theta}_2) = \sum_{k=1}^{n}(y_k - g(\mathbf{x}_k))^2 \qquad \text{Eq. 3}$$

where $g(\mathbf{x}_k)$ is the predicted value of the quantity of interest, for example k. The minimization of J was performed by the Levenberg–Marquardt algorithm. Being a second-order gradient optimization method, it requires the value of the gradient of the cost function with respect to the parameters, which was computed by the popular backpropagation algorithm (see, for example, Dreyfus[27]).

Model selection

As usual in the structural risk minimization framework,[25] models of increasing complexity were designed, and, for each complexity, the corresponding generalization ability was estimated. This can be achieved by various methods, including hold-out, cross-validation, leave-one-out and virtual leave-one-out. The latter method was used in the present study. It consists of estimating the generalization error of the model after training as:

$$E_p = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i}{1-h_{ii}}\right)^2} \qquad \text{Eq. 4}$$

where n is the number of training examples, r_i is the modeling error on observation i and h_{ii} is the leverage of observation i. The latter is the i-th diagonal element of matrix:

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}$$

where $\mathbf{Z}$ is the Jacobian matrix, whose element z_{ij} is given by $z_{ij} = \left(\partial g(\mathrm{x})/\partial\theta_j\right)_{\mathbf{x}=\mathbf{x}_i}$. Equation 4 is exactly equal to the leave-one-out estimation of the generalization error if the model is linear in its parameters (in that case E_p is called the PRESS – Predicted REsidual Sum of Squares – statistic), and it is approximate for non-linear models such as neural networks.[28]

Variable selection

Variable selection was performed by the random probe method, as described by Stoppiglia *et al.*[29] The principle of the method is the following: dummy candidate variables ('probes') are generated randomly, and appended to the set of 'true' candidate variables. All variables are ranked in order of decreasing relevance by the Gram–Schmitt orthogonalization method,[30] so that the relevance index of a candidate variable is its rank in that ranked list. The probe variables are obviously irrelevant, and the probability distribution function of their rank can be estimated. The rejection threshold is chosen such that the probability of selecting a variable that ranks below a probe variable (*i.e.* the probability of selecting a variable although it is probably irrelevant), has a predetermined value. More details on the random probe method, and alternative variable selection methods, can be found in Stoppiglia *et al.*[29] and Guyon *et al.*[31]

Estimation of confidence intervals for the prediction

Several approximate confidence intervals for the predictions of non-linear models have been proposed in the past.[32] In the present investigation, confidence intervals that involve the leverages (defined above) were used: the confidence interval for the prediction obtained for the vector of variables x, with confidence level α, is given by:

$$t_\alpha^{n-p}\, s\, \sqrt{\mathbf{z}^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{z}} \qquad \text{Eq. 5}$$

where t_α^{n-p} is a Student variable with n–p degrees of freedom, s is an estimate of the variance of the prediction error, and $\mathbf{z} = \partial g(\mathbf{x})/\partial \boldsymbol{\theta}$. The quantity under the square root sign is computed exactly as the leverages of the examples of the training set.

Software tools

The results described below were obtained with NeuroOne™ v.6 (a trademark of NETRAL S.A. <http://www.netral.com>), which implements the procedures described above for model training, variable selection, model selection and confidence interval estimation.[1]

Results

The results described in the present section illustrate various aspects of the predictive capabilities of the approach.

Prediction of glutathione concentrations

In order to unravel the relationship between the metabolism of glutathione and the concentrations of vitamins, trace elements, and proteins, the prediction of glutathione (GSH) was attempted. Table 1 shows the top of the ranked list of candidate variables, and the probability for each of them to be more relevant than a probe variable. The last two candidate variables were discarded by the random probe method (see above), leaving six selected variables. For simplicity, we first report results obtained on a small database of 57 patients (database I). In order to illustrate the influence of model complexity on prediction accuracy, Figure 2A shows the scatter plot (predicted value versus measured value) obtained on a training set by a model having three hidden neurons, and Figure 2B shows the scatter plot obtained with a more complex model (6 hidden neurons), trained on the same data.

Table 1 Variable selection for the prediction of glutathione levels in database I

Candidate variables	Probability for the candidate variable to be more relevant than a probe variable
Selenium	0.97
Protein thiol	0.97
Cu/Zn ratio	0.92
Vitamin E	0.83
Vitamin E/vitamin C ratio	0.72
Oxidized DNA	0.69
Vitamin C	0.49
Oxidized LDL	0.42

The top six variables were selected.

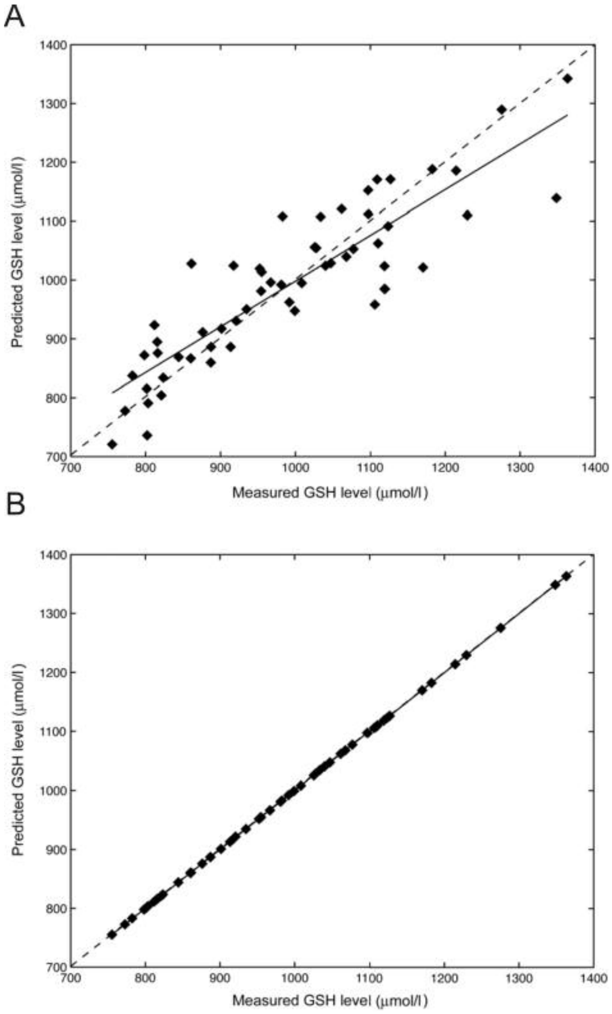

Figure 2 Prediction of the glutathione level (database I). (A) Scatter plot for a model with 3 hidden neurons (estimated generalization error 157 μmol/l). (B) Scatter plot for a model with 6 hidden neurons (estimated generalization error 24 μmol/l). Solid lines: linear regression lines of predictions versus observations

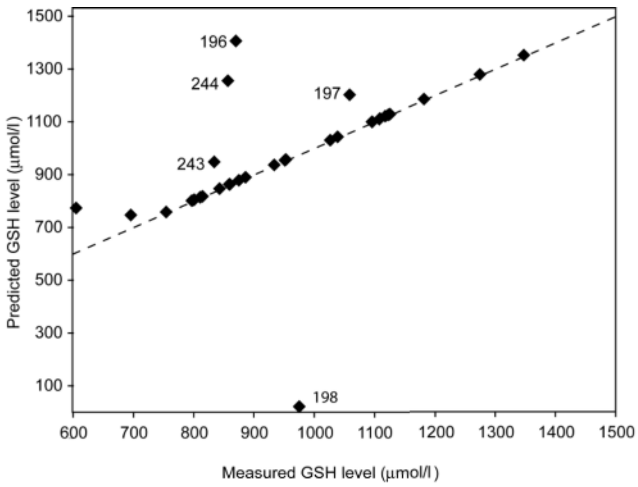

Figure 3 Prediction of glutathione levels on a test set (database I). Figures are the numbers of the corresponding records in the database

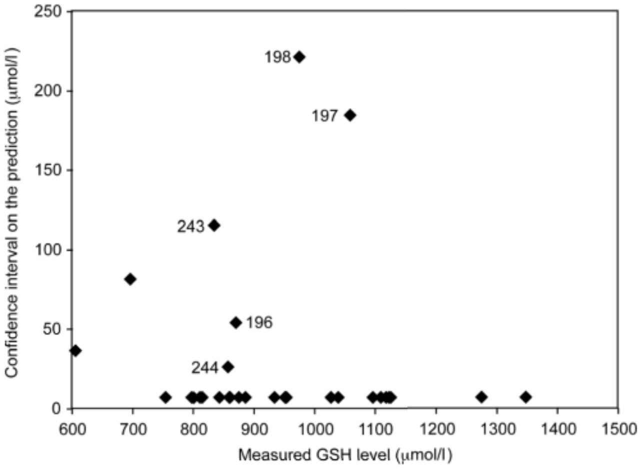

Figure 4 Confidence intervals for the prediction of glutathione concentrations (database I) by a model with 6 hidden neurons

The predictions of a model of intermediate complexity (4 hidden neurons) are shown in Figure 6. The estimated leave-one-out score for the three-hidden-neuron model is equal to 157 μmol/l, while it is equal to 24 μmol/l for the six-hidden-neuron model. The improvement, resulting from a controlled increase of the complexity of the model, is clearly apparent. The model with six hidden neurons, whose training results were most promising, was tested on fresh data (test set), *i.e.* on a set of examples that were used neither for training nor for variable and model selection. The results are shown in Figure 3. Clearly, most test examples are predicted as accurately as the training examples, with some exceptions:

1. Examples for which the measured glutathione concentration is lower than 750 μmol/l. Such examples lie below the concentration range in which training was performed (see Fig. 2): the prediction of these points cannot be expected to be accurate.
2. A few outliers. The figures printed by those points are the record numbers in the database; they are consecutive records, which gives strong suspicion of artefacts such as poor settings of the measurement apparatus on the day the analyses were performed, or data logging errors.

The estimations of the confidence intervals (Fig. 4) confirm that the predictions of those points should be granted low confidence: all predictions are assigned a small confidence interval, while the outliers have large confidence intervals. The importance of variable selection is illustrated in Figures 5 and 6. They show the scatter plots obtained for the prediction of glutathione concentration of the same training examples by models having the same number of hidden neurons, and, respectively, the three and six top variables of the ranked list (Table 1). As expected, the selection of relevant variables improves the quality of the prediction to a large extent. The above examples, obtained on a relatively small database, were intended to provide a striking illustration of the ability of the proposed approach to predict the glutathione concentration with satisfactory accuracy. The examples

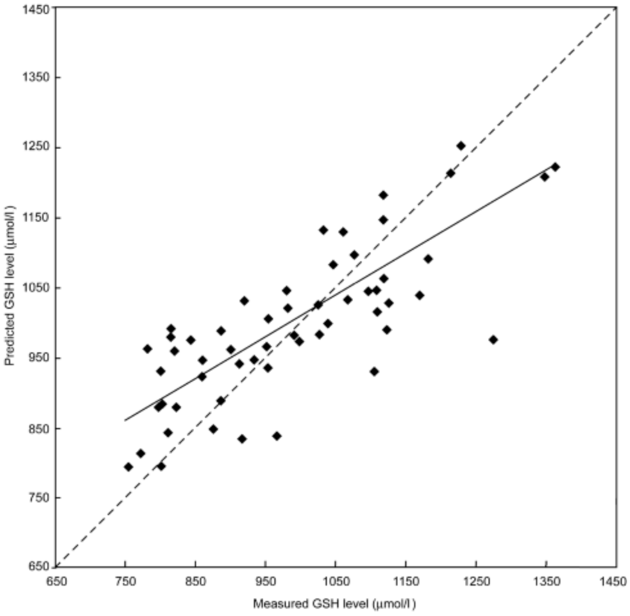

Figure 5 Prediction of glutathione concentration (database I) from 3 variables by a model with 4 hidden neurons. Estimated generalization error 175 μmol/l. Solid line: linear regression of the predicted values versus measured values

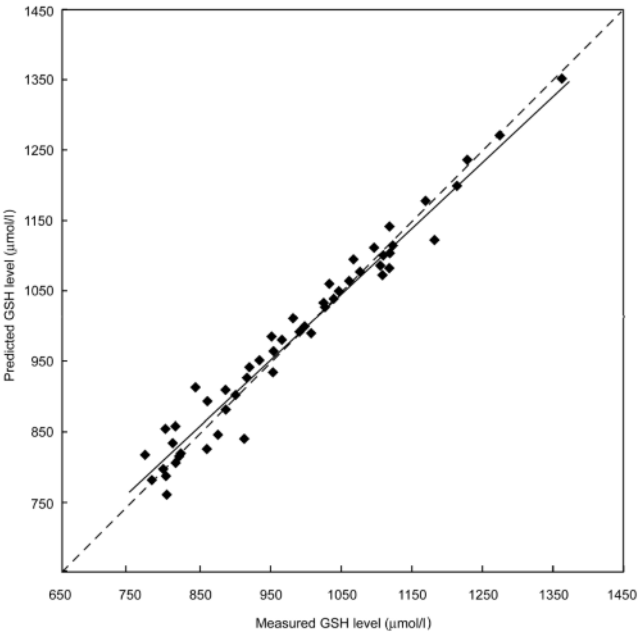

Figure 6 Prediction of glutathione concentration (database I) from 6 variables by a model with 4 hidden neurons. Estimated generalization error 153 µmol/l

described below show the predictive ability of models based on a larger database (database II), with larger inter-individual variability.

Prediction of glutathione and oxidized glutathione from exogenous antioxidants

In order to evaluate the relationship between vitamins, trace elements and proteins, glutathione concentration and the log ratio of glutathione to oxidized glutathione were predicted from the following selected concentrations (ranked in order of decreasing relevance): ratio Cu/Zn (93%), selenium (89%), protein thiol (89%), vitamin E (82%), ratio of vitamin C to vitamin E (63%). As indicated previously, the numbers in parentheses are the probability for the selected variable to be more relevant than a probe variable. Table 2 summarizes the selected variables of the models whose predictions are reported here and below. Figures 7 and 8 show that both quantities can be predicted with satisfactory accuracy. The estimated generalization errors are 174 µmol/l and 0.46 log units, respectively.

Prediction of robust biomarkers of oxidative stress: ratio 8-OH-dG/creatinine and oxidized LDL

The proposed approach allows the prediction of two robust biomarkers of oxidative stress: the ratios of the concentration of 8-OH-dG (8-hydroxy-2′-deoxyguanosine) to the concentration of creatinine, and the concentration of oxidized LDL (low density lipoproteins). The results are shown in Figures 9 and 10. For the 8-OH-dG/creatinine concentration ratio, the selected variables (Table 2) were the Cu/Zn concentration ratio (98%), the glutathione to oxidized glutathione concentration ratio (98%), and the concentrations of vitamin E (90%), selenium (84%), vitamin C (75%) and protein thiol (57%). The estimated generalization error was 8.9. For the prediction of oxidized LDL, the log of the concentration (µmol/l) was predicted, because of the large range of measured concentrations. The selected variables were protein thiol (99%), vitamin E (99%),

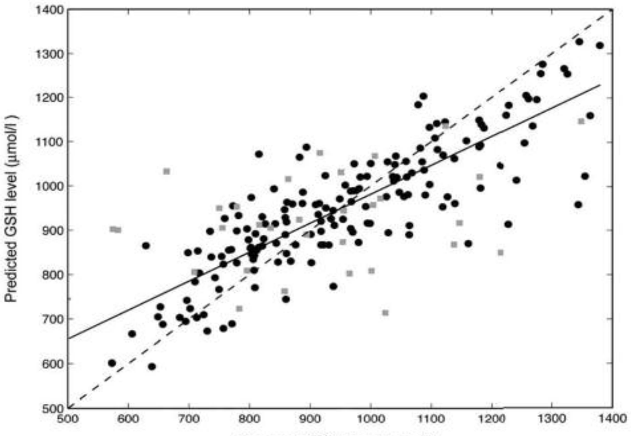

Figure 7 Prediction of glutathione concentration from exogenous antioxidants (database II, 208 examples). Dots, training set; gray squares, test set. Solid line, linear regression of predicted values versus measured values

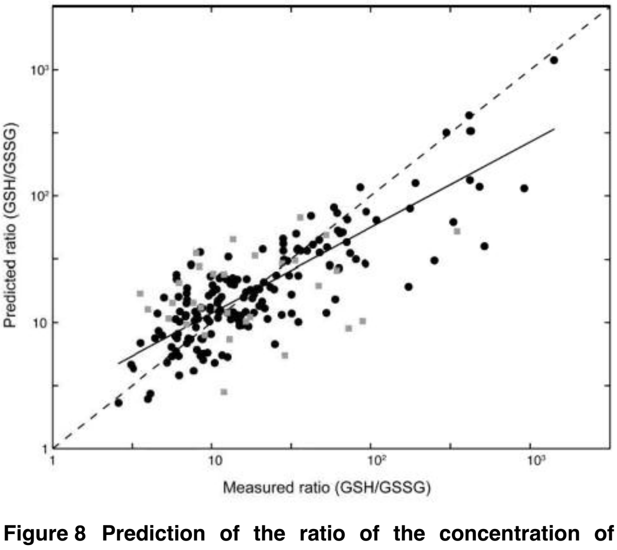

Figure 8 Prediction of the ratio of the concentration of glutathione to the concentration of oxidized glutathione, from exogenous antioxidants (database II, 208 examples). Dots, training set; gray squares, test set. Solid line, linear regression of predicted values versus measured values

Table 2 Variable selection for the predictions performed on database II

	Predicted quantity			
Candidate variable	Glutathione concentration	Glutathione to oxidized glutathione concentration ratio	8-OH-dG to creatinine concentration ratio	Oxidized LDL concentration
Cu/Zn ratio	0.93	0.93	0.98	0.81
Protein thiol	0.89	0.89	0.57	0.99
Selenium	0.89	0.89	0.84	0.89
Vitamin C	NS	NS	0.75	0.98
Vitamin E	0.82	0.82	0.90	NS
Vitamin E/vitamin C ratio	0.63	0.63	NS	0.99
Glutathione/oxidized glutathione	NS	NS	0.98	0.94
Glutathione	NS	NS	NS	0.81

Numbers are the probability for the candidate variable to be more relevant than a probe variable. NS, not selected.

vitamin C (98%), GSSG (94%), selenium (89%), GSH (81%), and Cu/Zn concentration ratio. The estimated generalization error was 0.22 log units.

Discussion

For the first time, this study has validated the feasibility of predicting concentrations of biomarkers of oxidative stress based on measurements of exogenous and endogenous antioxidants in plasma and urine from a large clinical and biological database derived from patients presenting a wide range of clinical disorders involving chronic inflammation and oxidative stress. Distinct profiles of biomarkers of oxidative stress can be described depending on the nature of the oxidative insult. We addressed the question of the choice of pertinent oxidative stress biomarkers in the context of chronic inflammatory disease, and our data highlight three clusters of biomarkers: exogenous (vitamins E and C, copper, zinc, selenium, thiols) and endogenous (GSH and GSSG) biomarkers of antioxidant status, and terminal biomarkers of oxidative damage (oxidized LDL and 8-OHdG). Importantly, as oxidised biomolecules are typically replaced by their native counterparts so as to diminish the impact of oxidative damage *in vivo*, circulating levels of biomarkers of oxidative damage are often difficult to measure, largely as a result of their low levels and of the high analytical sensitivity required. The innovative application of a machine-learning approach to the prediction of oxidative stress allows us for the first time to predict abnormalities in biomarkers of one group, primarily those in terminal biomarkers of oxidative damage, relative to biomarker abnormalities

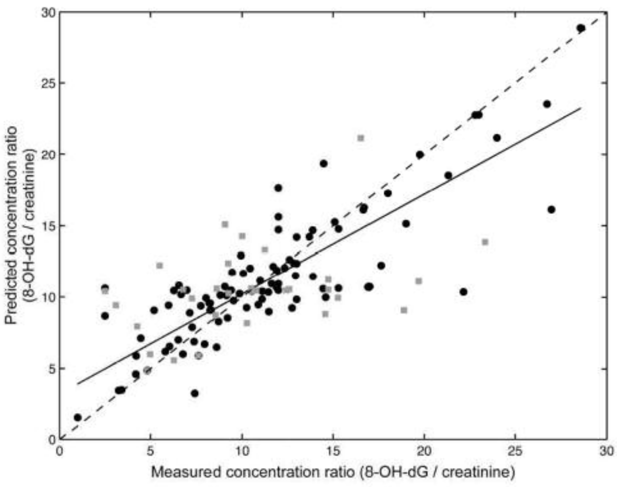

Figure 9 Prediction of the concentration ratio of 8-OH-dG to creatinine (database II, 131 examples). Dots, training set; gray squares, test set. Solid line, linear regression of predicted values versus measured values

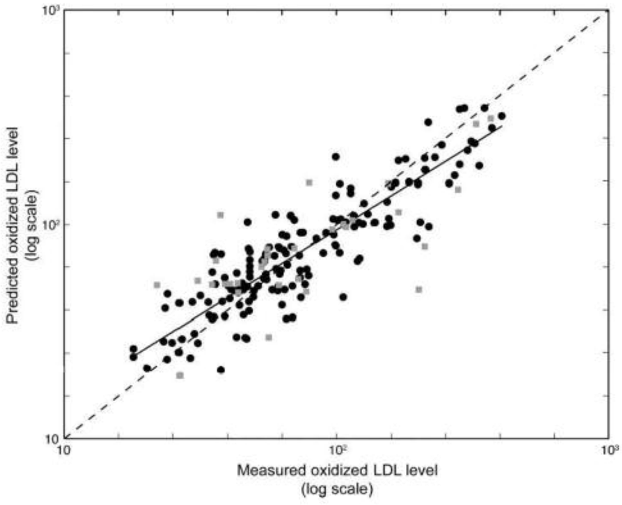

Figure 10 Prediction of the concentration of oxidized LDL (database II, 197 examples). Dots, training set; gray squares, test set. Solid line, linear regression of predicted values versus measured values

of another group. The biomarkers of oxidative damage assayed in this study (*i.e.* circulating oxidized LDL as a marker of lipid peroxidation, and urine 8OHdG/creatinine as a marker of DNA oxidation) could be predicted by the biomarkers of antioxidant status, especially Cu/Zn concentration ratio, GSH/GSSG ratio, and concentrations of vitamins E and C and selenium. Levels of endogenous antioxidants, such as glutathione, could thus be predicted on the basis of levels of exogenous antioxidants (antioxidative vitamins and some trace elements), and preferentially using concentrations of selenium, total thiols, copper/zinc ratio, vitamin E concentration, and finally the vitamin C/vitamin E ratio (Figs 2 and 3). The predictive power of this approach progressively diminishes, however, when the number of biomarkers decreases from 6 to 3 (Figs 5 and 6), thereby implying that a minimal number of associated biomarkers is required for the reliable prediction of oxidative stress (Fig. 7). Another biomarker related to plasma levels of endogenous antioxidants, the GSH/GSSG ratio, could equally be predicted under similar conditions, with the exception of the order of predictive power of other biomarkers (order of relevance: the copper/zinc ratio, concentrations of selenium, thiols, vitamin E, and finally the vitamin C/vitamin E ratio). The possibility of predicting biomarkers of oxidative damage assayed in this study by biomarkers of antioxidant status indicates that the pertinence level attained leads to a more appropriate choice of oxidative stress biomarkers, and the predictive power allows reduction in the number of biomarkers to be evaluated, thereby resulting in greater technical and economical feasibility. Indeed, the appropriate choice of biomarkers is essential for an informative and pertinent diagnostic approach. This choice constitutes a critical feature of clinical studies involving antioxidant supplementation as it provides key information on the efficacy of the therapeutic response. The absence of data on baseline levels of oxidative stress may have resulted in antioxidant supplementation in subjects displaying normal levels of oxidative stress biomarkers, and who would not be predicted to display further benefit. The direct relationship between the efficacy of antioxidant supplementation and baseline levels of oxidative stress has been recently demonstrated.[33] The inability of dietary antioxidants to reduce the incidence of cardiovascular disease or cancer could thus result from the lack of knowledge of baseline levels of oxidative stress in the populations studied.[13–15] Our approach, which involves extensive characterisation of the level of oxidative stress in a given subject using a set of relevant biomarkers whose interrelationships are well understood and can be mathematically predicted, may open new horizons in the routine assessment of oxidative stress in a clinical setting. It is important to mention that our studies reveal the critical role of methodological tools including model selection, variable selection, and confidence interval estimation. From the machine learning point of view, the main open question is the following: are the present results optimal or can they be improved, for example, by using different learning machines, or by implementing regularization as in support vector machines, or by designing ‘committees of machines’? That question can be answered if, and only if, an estimate of the experimental error is available: if the uncertainty of the prediction is of the order of magnitude of the uncertainty in the measurement, no improvement can be expected. If the experimental uncertainty is substantially lower than the prediction error, however, then the results can be improved. The present informative findings on the prediction of levels of oxidative damage biomarkers using the measured levels of exogenous and endogenous antioxidants in a French population reveal that it is worthwhile pursuing this study on a large set of biological samples derived from patient populations of distinct ethnicity, life-style and diet. The patient populations which will be targeted should include a wide spectrum of chronic diseases involving chronic inflammation and oxidative stress in order to allow further evaluation of the present innovative approach.

References

1. Halliwell B, Gutteridge JM. *Free Radicals in Biology and Medicine*. Oxford: Clarendon, 1999.
2. Van Lenten BJ, Navab M, Shih D, Fogelman AM, Lusis AJ. The role of high-density lipoproteins in oxidation and inflammation. *Trends Cardiovasc Med* 2001; 11: 155–161.
3. Stocker R, Keaney Jr JF. Role of oxidative modifications in atherosclerosis. *Physiol Rev* 2004; 84: 1381–1478.
4. Barnham KJ, Masters CL, Bush AI. Neurodegenerative diseases and oxidative stress. *Nat Rev Drug Discov* 2004; 3: 205–214.
5. Floyd RA, Hensley K. Oxidative stress in brain aging. Implications for therapeutics of neurodegenerative diseases. *Neurobiol Aging* 2002; 23: 795–807.
6. Clarke R, Armitage J. Antioxidant vitamins and risk of cardiovascular disease. Review of large-scale randomised trials. *Cardiovasc Drugs Ther* 2002; 16: 411–415.
7. Coulter ID, Hardy ML, Morton SC *et al.* Antioxidants vitamin C and vitamin E for the prevention and treatment of cancer. *J Gen Intern Med* 2006; 21: 735–744.
8. Pham DQ, Plakogiannis R. Vitamin E supplementation in cardiovascular disease and cancer prevention: Part 1. *Ann Pharmacother* 2005; 39: 1870–1878.
9. Stanner SA, Hughes J, Kelly CN, Buttriss J. A review of the

epidemiological evidence for the 'antioxidant hypothesis'. *Public Health Nutr* 2004; 7: 407–422.

10. Dutta A, Dutta SK. Vitamin E and its role in the prevention of atherosclerosis and carcinogenesis: a review. *J Am Coll Nutr* 2003; 22: 258–268.
11. Padayatty SJ, Katz A, Wang Y *et al.* Vitamin C as an antioxidant: evaluation of its role in disease prevention. *J Am Coll Nutr* 2003; 22: 18–35.
12. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis. *JAMA* 2007; 297: 842–857.
13. Witztum JL, Steinberg D. The oxidative modification hypothesis of atherosclerosis: does it hold for humans? *Trends Cardiovasc Med* 2001; 11: 93–102.
14. Heinecke JW. Is the emperor wearing clothes? Clinical trials of vitamin E and the LDL oxidation hypothesis. *Arterioscler Thromb Vasc Biol* 2001; 21: 1261–1264.
15. Hercberg S, Galan P, Preziosi P *et al.* The SU.VI.MAX study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. *Arch Intern Med* 2004; 164: 2335–2342.
16. Therond P, Bonnefont-Rousselot D, Davit-Spraul A, Conti M, Legrand A. Biomarkers of oxidative stress: an analytical approach. *Curr Opin Clin Nutr Metab Care* 2000; 3: 373–384.
17. Dalle-Donne I, Rossi R, Colombo R, Giustarini D, Milzani A. Biomarkers of oxidative damage in human disease. *Clin Chem* 2006; 52: 601–623.
18. Kadiiska MB, Gladen BC, Baird DD *et al.* Biomarkers of oxidative stress study II: are oxidation products of lipids, proteins, and DNA markers of CCl_4 poisoning? *Free Radic Biol Med* 2005; 38: 698–710.
19. Kadiiska MB, Gladen BC, Baird DD *et al.* Biomarkers of oxidative stress study III. Effects of the nonsteroidal antiinflammatory agents indomethacin and meclofenamic acid on measurements of oxidative products of lipids in CCl_4 poisoning. *Free Radic Biol Med* 2005; 38: 711–718.
20. Omaye ST, Turnbull JD, Sauberlich HE. Selected methods for the determination of ascorbic acid in animal cells, tissues, and fluids. *Methods Enzymol* 1979; 62: 3–11.
21. Zhao B, Tham SY, Lu J, Lai MH, Lee LK, Moochhala SM. Simultaneous determination of vitamins C, E and beta-carotene in human plasma by high-performance liquid chromatography with photodiode-array detection. *J Pharm Sci* 2004; 7: 200–204.
22. Sturup S, Hayes RB, Peters U. Development and application of a simple routine method for the determination of selenium in serum by octopole reaction system ICPMS. *Anal Bioanal Chem* 2005; 381: 686–694.
23. Tietze F. Enzymic method for quantitative determination of nanogram amounts of total and oxidized glutathione: applications to mammalian blood and other tissues. *Anal Biochem* 1969; 27: 502–522.
24. Ellman GL. Tissue sulfhydryl groups. *Arch Biochem Biophys* 1959; 82: 70–77.
25. Vapnik V. *The Nature of Statistical Learning Theory*. Berlin: Springer, 1999.
26. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford: Clarendon, 1995.
27. Dreyfus G. *Neural Networks, Methodology and Applications*. Berlin: Springer, 2005.
28. Monari G, Dreyfus G. Local overfitting control via leverages. *Neural Comput* 2002; 14: 1481–1506.
29. Stoppiglia H, Dreyfus G, Dubois R, Oussar Y. Ranking a random feature for variable and feature selection. *J Machine Learning Res* 2003; 3: 1399–1414.
30. Chen S, Billings SA, Luo W. Orthogonal least squares methods and their application to non-linear system identification. *Int J Control* 1989; 50: 1873–1896.
31. Guyon I, Gunn S, Nikravesh M, Zadeh LA. *Feature Extraction: Foundations and Applications*. Berlin: Springer, 2006.
32. Bates DB, Watts DG. *Nonlinear Regression Analysis and its Applications Wiley Series in Probability and Mathematical Statistics*. New York: John Wiley, 1988.
33. Block G, Jensen CD, Morrow JD *et al.* The effect of vitamins C and E on biomarkers of oxidative stress depends on baseline level. *Free Radic Biol Med* 2008; 45: 377–384.