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

Alban Magon de La Villehuchet, Michel Brack, Gerard Dreyfus, Yacine Oussar, Dominique Bonnefont-Rousselot, Anatol Kontush, M John Chapman

A MACHINE LEARNING APPROACH TO THE PREDICTION OF OXIDATIVE STRESS IN CHRONIC INFLAMMATORY DISEASE

A. Magon de la Villehuchet[a], M. Brack[b] , G. Dreyfus[a]*, Y. Oussar[a], D. Bonnefont-Rousselot[c], A. Kontush[d] and M.J. Chapman[d]

[a] École Supérieure de Physique et de Chimie Industrielles, ESPCI - Paristech, Laboratoire d'Électronique (CNRS UMR 7084), 10 rue Vauquelin, 75005 Paris, France

[b] Clinical Center for Oxidative Stress, rue Marbeuf, 75008 Paris, France

c Department of Biochemistry, EA 3617, Faculty of Pharmacy, Paris5 University, Paris, France

[d] Dyslipoproteinemia and Atherosclerosis Research Unit (U.551), National Institute for Health and Medical Research (INSERM), Hôpital de la Pitié, Paris, France

[*]**Corresponding author:** G. Dreyfus, ESPCI, laboratoire d'Électronique, 75005 Paris, France. Tel. +33 1 40 79 45 41 . Fax +33 1 47 07 13 93. E-mail Gerard.Dreyfus@espci.fr

Running title: Machine learning approach to predict oxidative stress

Abstract

Oxidative stress is involved 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 predict levels of oxidative stress in human subjects. From a database of biochemical analyses of oxidative stress biomarkers in blood, plasma and urine, nonlinear models have been designed, with a statistical methodology that includes variable selection, model training and model selection. We show 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/or endogenous antioxidants.

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

1. 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 in subjects with elevated levels of oxidative stress, the inability of dietary antioxidants to reduce the incidence of 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 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 principal among 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 oxidised 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 which occur in vivo result in distinct profiles of biomarkers of oxidative stress. The diversity of oxidative species implies that the choice of biomarkers which 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 common oxidative pathways; 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 large 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 ability of our models.

2. Clinical database: contents and protocols

Subjects. The Clinical Center 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 equally measured in each subject. In 150 subjects, a second assessment of systemic oxidative stress followed within 4 to 6 months after the first visit. Majority of subjects presented with 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, 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 Center in the absence of any known symptoms and who were free of a clinical diagnosis were considered as healthy controls. Rest of the subjects (n=129) presented with relatively rare pathological conditions ($n < 20$ for each specific disease) and was 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 minutes. 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 24h.

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), endogenous antioxidants (reduced and oxidised glutathion, 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 oxidised glutathione. Reduced (GSH) and oxidized (GSSG) glutathione were measured in whole blood using a Bioxytech GSH/GSSG-412TM kit (OxisResearch, Portland, 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, England).

Determination of oxLDL. Levels of oxLDL were measured using a competitive enzyme-linked immunosorbent assay (ELISA) kit supplied by Immunodiagnostik (Germany; inter- and intraassay 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 endpoint 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, Austria) using Cu^{2+}-oxidized LDL as an antigen (inter- and intraassay coefficients of variation, 10.5 %).

Determination of lipid peroxides. Lipid peroxides were assessed in plasma using an Oxystat spectrophotometric kit (Biomedica, Vienna, Austria) 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 and expressed as ng/mg creatinine.

3. The statistical machine learning approach: building models by training from examples

3.1. A cursory 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 understood as a physical object, but as a set of algorithms and procedures that are implemented on a computer. The mathematical foundations of statistical machine learning are described in [25]. In the present article, we focus on the application of machine learning to the design of predictive models in the form of nonlinear parameterized functions. In other words, functions are derived that

- “explain”, in a statistical sense, the existing values,
- can generalize to hitherto unknown situations, i.e. can predict the outcome of future measurements.

3.1.1. Training

Training is an algorithmic procedure whereby the parameters of the model are adjusted in order to fit the measurements present in a database called “training set”. There is a wide variety of training algorithms, depending on the task to be fulfilled and on the learning machine to be trained. The specific procedure used in the present study is described cursorily in section 3.2.3.

3.1.2. Model selection and the bias-variance dilemma

Model selection is a central task in statistical machine learning. It involves solving the so-called “bias-variance dilemma”, i.e. defining the appropriate complexity of the model, given the training data. The complexity of the model can be characterized roughly by the number of its adjustable parameters. If a model is insufficiently complex, it is unable to learn the training data; conversely, if the model is too complex, it adjusts very accurately to the training data, thus to the noise present in it (a phenomenon known as *overfitting*), and generalizes poorly. A model that is too simple has high bias (it reproduces the training data inaccurately), but low variance (it is insensitive to the details of the training data), while an overly complex model has low bias (it learns the training data accurately) but high variance (being highly sensitive to

the noise present in the training data, it generalizes poorly). Metaphorically, the "intelligence" of a model results from a tradeoff between ignorance (low complexity, inability to learn) and stupidity (excessive complexity, inability to generalize), as illustrated on Figure 1. The estimation of the generalization ability of models is a basic ingredient in the model selection procedure. The model selection method used in the present study is described in section 3.2.4.

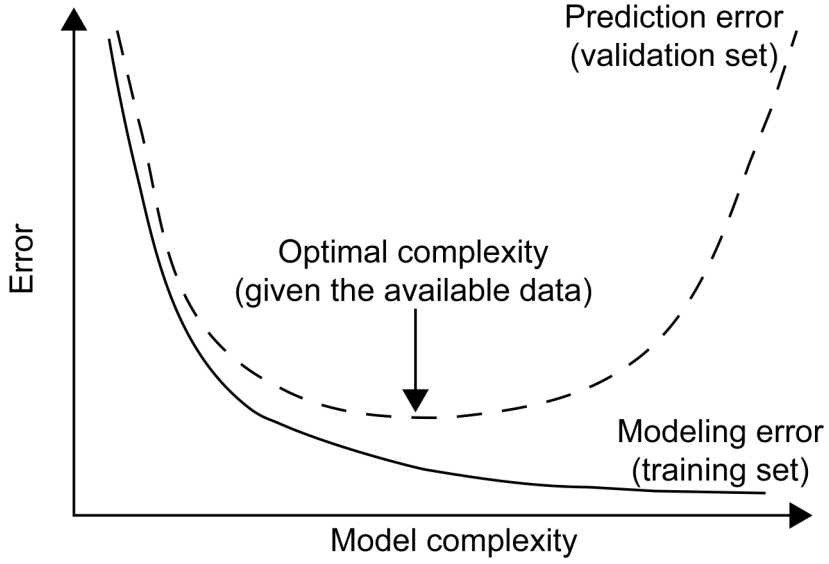


Figure 1
Pictorial representation of the bias-variance tradeoff.

3.1.3. Variable selection

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 modeled 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.

On a statistical basis, variable selection involves the following approach. It is assumed that the relevance of the candidate variables is estimated by the value of an appropriately defined

"relevance index": the larger the value of that index for a given variable, the more relevant the variable. It can then be expected that the probability distribution functions of the relevance index for relevant variables and for irrelevant variables will have little, or no, overlap, as shown in Figure 2. Variable selection consists of finding a decision threshold, so that all candidate variables with relevance index above the threshold will be retained, while all candidate variables with relevance index below the threshold will be discarded. The variable selection method used in the present study is described in section 3.2.2.

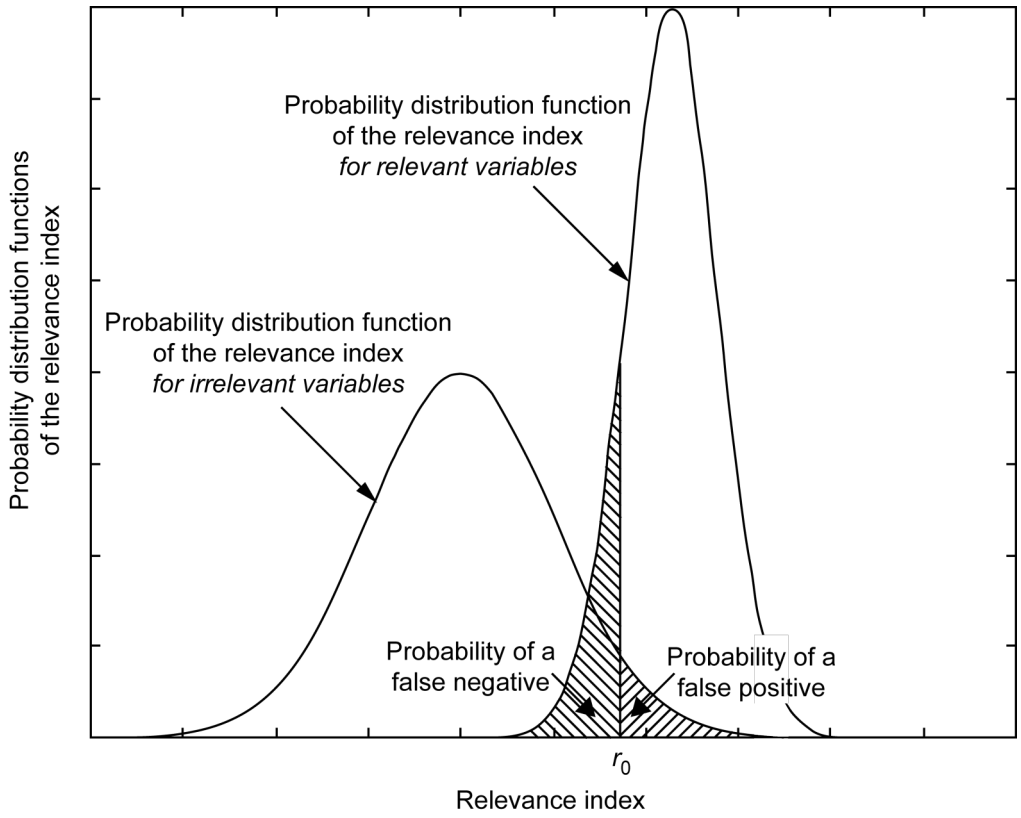


Figure 2
Statistical description of the variable selection problem; r_0 is a decision threshold. A candidate variable whose relevance index is smaller than r_0 will be considered irrelevant, hence discarded. The hatched areas are the probability of a false positive (retaining a candidate variable although it is actually irrelevant) and of a false negative (discarding a candidate variable although it is actually relevant).

3.1.4. Estimation of confidence intervals for the prediction

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, …)

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 the parameters have no specific meaning; therefore, 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 in section 3.2.5.

3.2. Model design

This section describes the specific methodology used to derive the results described in section 4.

3.2.1. The learning machines: "neural networks"

Neural networks are learning machines that were primarily intended to model brain functions, but turned out to have useful properties in their own right, unrelated to their "biological" origin; for introductory textbooks, see for instance [26] and [27]. A neuron is a nonlinear, bounded, parameterized function. The neural networks used in the present study are linear combinations of so-called "hidden" neurons; such neural networks are termed "feedforward neural networks" or "multilayer Perceptrons".

More specifically, in the present study, a neuron performs an s-shaped ("sigmoid"), function of a linear combination of its variables. The neuron computes the value of f defined as:

$$f = \tanh(\boldsymbol{\theta} \cdot \mathbf{x}) \tag{1}$$

where $\boldsymbol{\theta}$ is the vector of parameters (sometimes called "synaptic weights" in the literature) 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 "feedforward 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 **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:

$$g(\mathbf{x}) = [illegible] \cdot \mathbf{f}([illegible]\mathbf{x}). \qquad (2)$$

Feedforward neural networks are frequently described pictorially as shown on Figure 3.

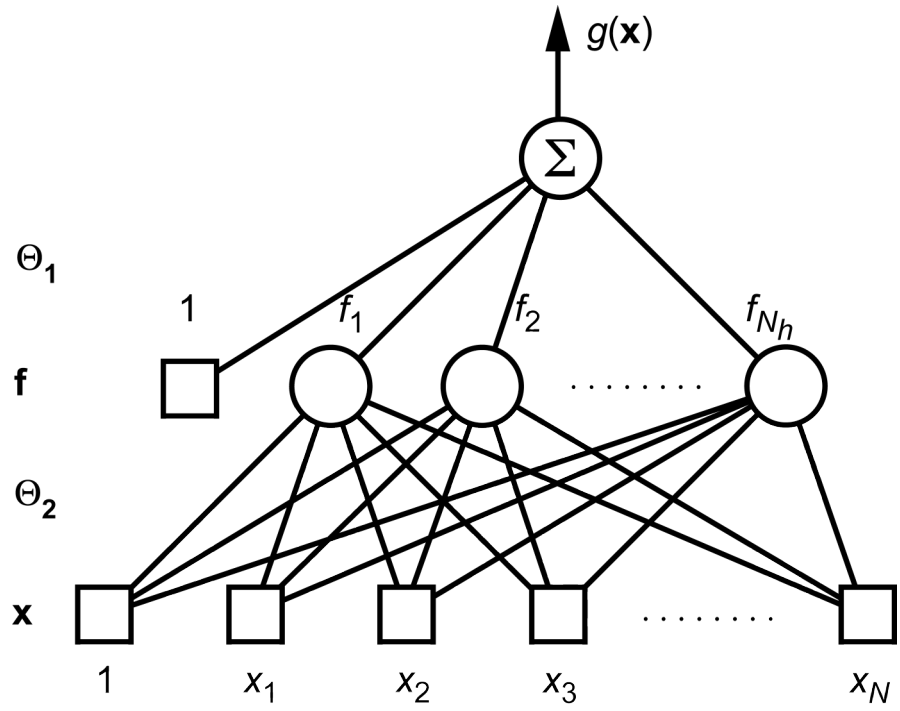


Figure 3
A feedforward neural network with N variables and N_h hidden neurons.

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

Neural networks are *parsimonious*: it is clear from relation (2) that the model $g(\mathbf{x})$ is nonlinear with respect to the parameters of matrix Θ_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 (as described in section 3.1.2) than polynomial models and, more generally, than linear-in-their-parameters models.

3.2.2. Variable selection

Variable selection was performed by the random probe method, as described in [28]. 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 [29], 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 threshold is chosen such that the probability of selecting a variable that ranks below a probe variable, has a predetermined value. If there is a wealth of data, the threshold can be set relatively high, because one can afford to retain a variable although it is irrelevant; conversely, if data is sparse, the threshold is set to a low value, so as to keep the probability of a false positive low. More details on the random probe method, and alternative variable selection methods, can be found in [28] and [30].

3.2.3. Training

Assume that a data base (called "training set") is available; it contains n examples, i.e. n pairs $\{\mathbf{x}_k, y_k\}$, where $\mathbf{x}_k$ is the vector of selected variables for example k, and y_k is the corresponding measured value of the quantity of interest. During training, the parameters of the network are

adjusted so as to minimize the least squares cost function, i.e. the sum of squared modeling errors on the examples present in the training set:

$$J(\blacksquare\blacksquare) = \sum_{k=1}^{n} \left(y_k - g(\mathbf{x}_k) \right)^2 \tag{3}$$

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 e.g. [27]).

3.2.4. Model selection

As usual in the *empirical risk minimization* framework [25], models of increasing complexity were designed, and, for each degree of complexity, the corresponding generalization ability was estimated. This can be achieved in various ways, including hold-out, cross-validation, and leave-one-out. The latter method involves withdrawing an example from the data set, train the learning machine on the n-1 other examples, compute the prediction error on that example, reinsert the left-out example into the database, and iterate the whole procedure n times. The *leave-one-out score* E is computed as:

$$E = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(r_i^{-i} \right)^2} \tag{4}$$

where r_i^{-i} is the prediction error on example i when it is withdrawn from the training set. The leave-one-out score is proved to be an unbiased estimate of the generalization error [25]. Hence, that procedure is accurate, but very computer-intensive. It can be used only for small data sets. For large data sets, D-fold cross-validation is applicable: instead of withdrawing a single example from the data set, a fraction 1/D of the data set is withdrawn, training is

performed on the remaining examples, and the procedure is iterated D times (typically $D = 5$), so that each example is used for validation once and only once.

For medium-sized data sets (a few tens to a few hundreds examples), as is the case in the present study, *virtual leave-one-out* is an attractive procedure [31], which provides an approximation of the leave-one-out score of nonlinear models, and an exact evaluation of the leave-one-out score for linear models (in which case it is called the PRESS – Predicted Residual Sum of Squares – statistic). Virtual leave-one-out consists of training the model on the whole data set, and approximating the modeling error that would have occurred on example i if it had been withdrawn from the training set as:

$$r_i^{-i} \approx \frac{r_i}{1 - h_{ii}} \tag{5}$$

where r_i is the actual modeling error on example i, and h_{ii} is the i-th diagonal element of the "hat matrix" $\mathbf{H}$:

$$\mathbf{H} = \mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}. \tag{6}$$

$\mathbf{Z}$ is the Jacobian matrix, whose element z_{ij} is given by $z_{ij} = \left(\frac{\partial g(\mathbf{x})}{\partial \theta_j}\right)_{\mathbf{x}=\mathbf{x}_i}$. Relation (5) is exact for models that are linear in their parameters, and approximate otherwise.

In analogy to relation (4), the virtual leave-one-out score E_p is defined as:

$$E_p = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{r_i}{1 - h_{ii}}\right)^2}. \tag{7}$$

h_{ii} is called the *leverage* of example i, because it reflects the influence of example i on the model [31]. The computation of $\mathbf{H}$ is straightforward, so that virtual leave-one-out is essentially n times as fast as the original leave-one-out procedure.

3.2.5. Estimation of confidence intervals for the prediction

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

$$t_{\alpha}^{n-p} s \sqrt{\mathbf{z}^{\mathrm{T}} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{z}} \tag{8}$$

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 $\mathbf{z} = \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$. The quantity under the square root sign is computed exactly as the leverages of the examples of the training set.

3.3. Software tools

The results described below were obtained with NeuroOne™ v.6, which implements the procedures described above for model training, variable selection, model selection and confidence interval estimation[1].

4. Results

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

4.1. Prediction of glutathione concentrations

In order to unravel the relationship between the metabolism of glutathione and the concentrations of vitamins, oligo-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

[1] NeuroOne™ is a trade mark of NETRAL S.A. (http://www.netral.com)

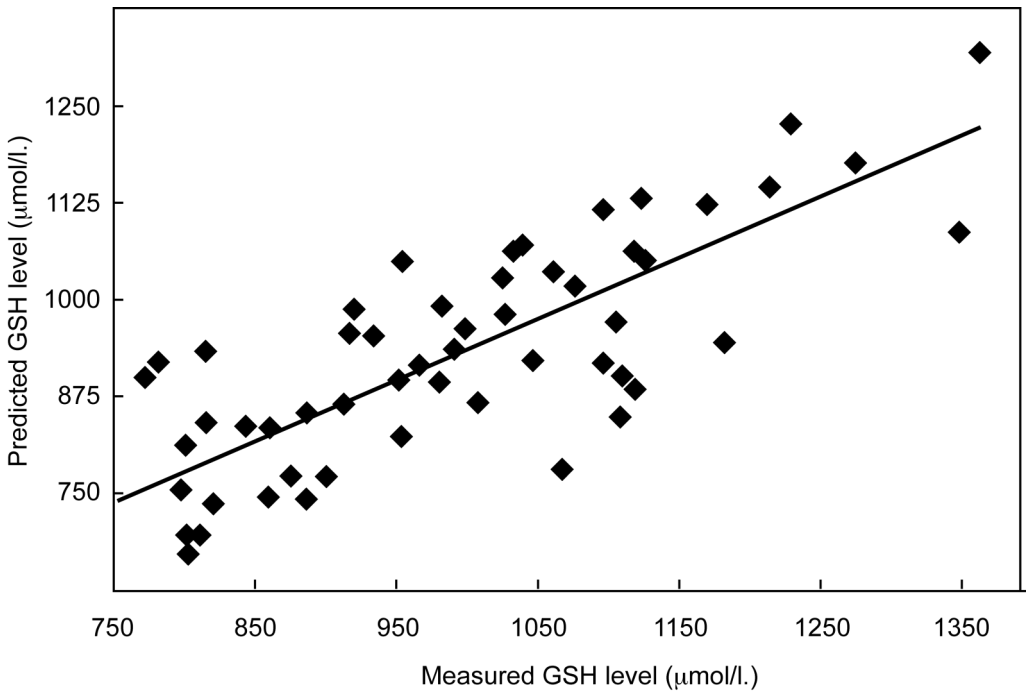
variables were discarded by the random probe method (see section 3.1.3), leaving six selected variables.

Candidate variables	*Probability of the candidate variable being 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

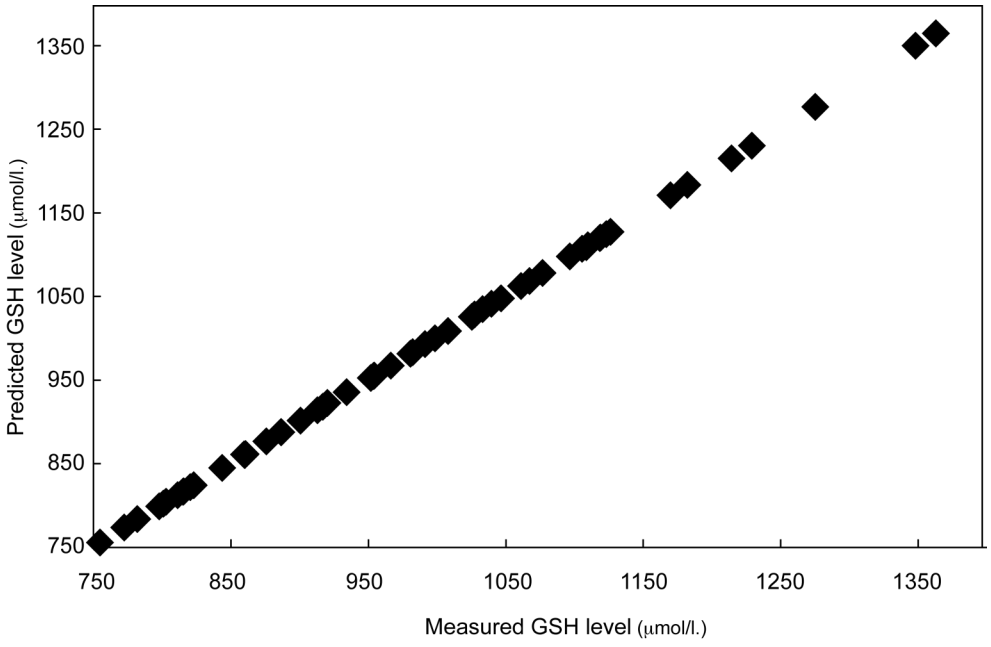
Table 1

Variable selection for the prediction of glutathione; the top six variables were selected.

For simplicity, we first report results obtained on a small database of 57 patients. In order to illustrate the influence of model complexity on prediction accuracy, Figure 4a shows the scatter plot (predicted value versus measured value) obtained with a model having three hidden neurons, and Figure 4b shows the scatter plot obtained with a more complex model (6 hidden neurons), trained on the same data. The predictions of a model of intermediate complexity (4 hidden neurons) are shown in Figure 8. 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.



(a)



(b)

Figure 4
(a) scatter plot for a model with 3 hidden neurons; (b) scatter plot for a model with 6 hidden neurons.

The model with six hidden neurons 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 on Figure 5.

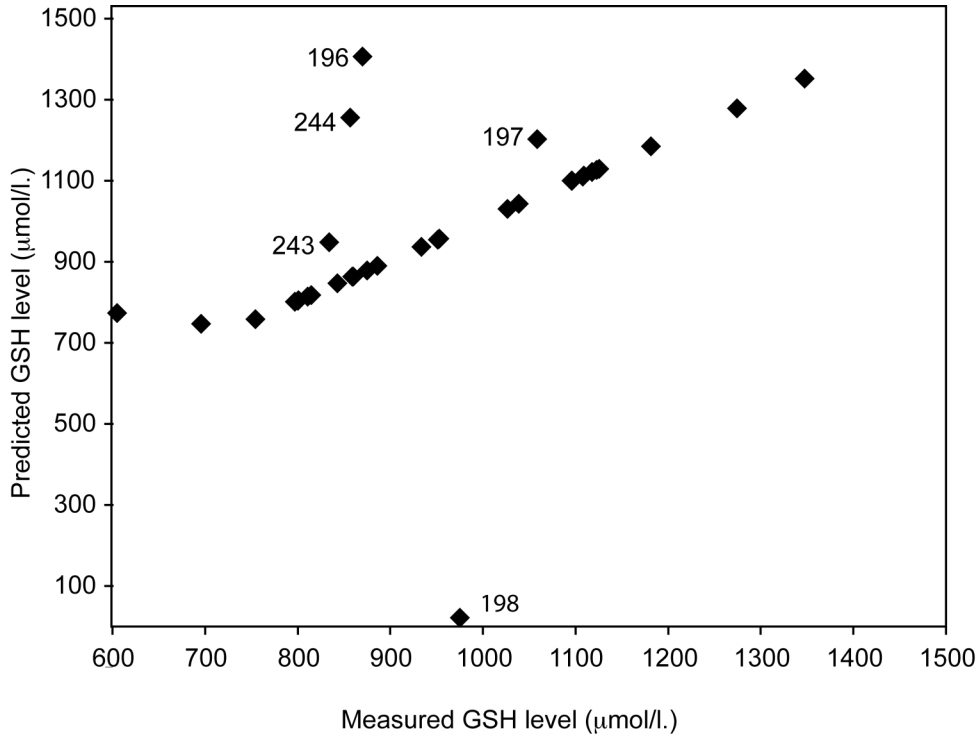


Figure 5
Prediction of GSH on a test set. Figures are the numbers of the corresponding records in the database.

Clearly, most examples are predicted accurately, with some exceptions:

- Examples for which the measured glutathione concentration is lower than 750 μmol/l. Those examples lie below the concentration range in which training was performed (see Figure 4): the prediction of such points cannot be expected to be accurate;
- 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 artifacts 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, reported on Figure 6, confirm that the predictions of those points should be accorded low confidence: all predictions are assigned a small confidence interval, while the outliers have large confidence intervals.

The importance of variable selection is illustrated on Figure 7 and Figure 8. They show the scatter plots obtained for the prediction of glutathione concentration by models having the same complexity (4 hidden neurons), and, respectively, the three and six top variables of the ranked list (Table 1). As expected, the incidence of relevant variables improves the quality of the prediction to a large extent.

The above examples, obtained on a relatively small database, illustrate clearly the ability of the proposed approach to predict the glutathione concentration with satisfactory accuracy.

The examples described in the next section show the predictive ability of models based on a larger database (200 patients), with larger inter-individual variability.

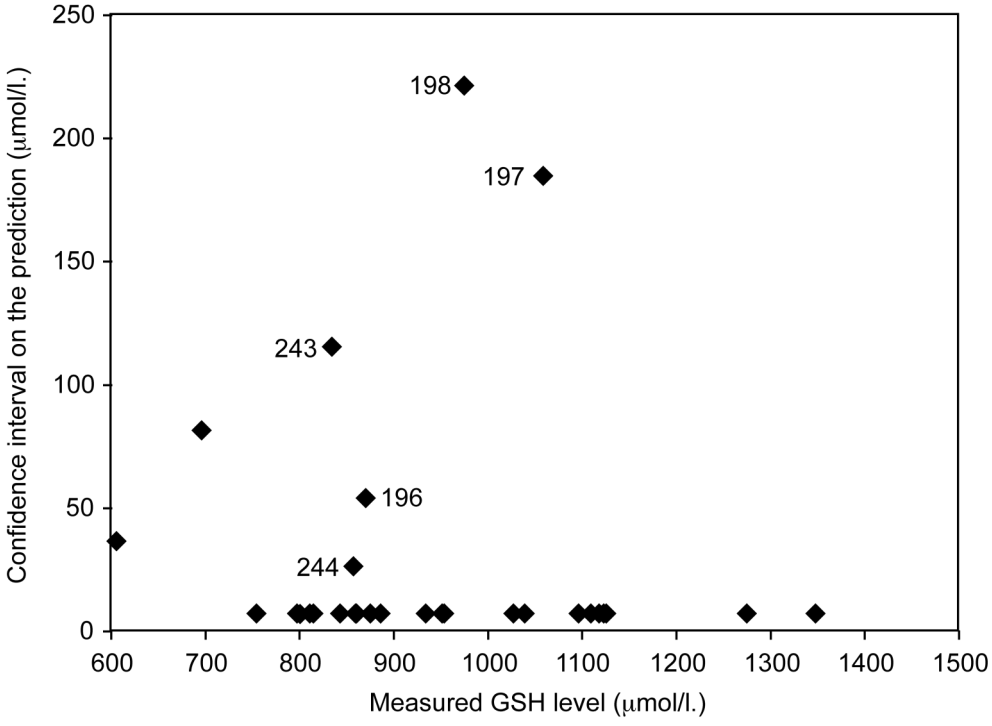


Figure 6
Confidence intervals on the predictions of a model with 6 hidden neurons

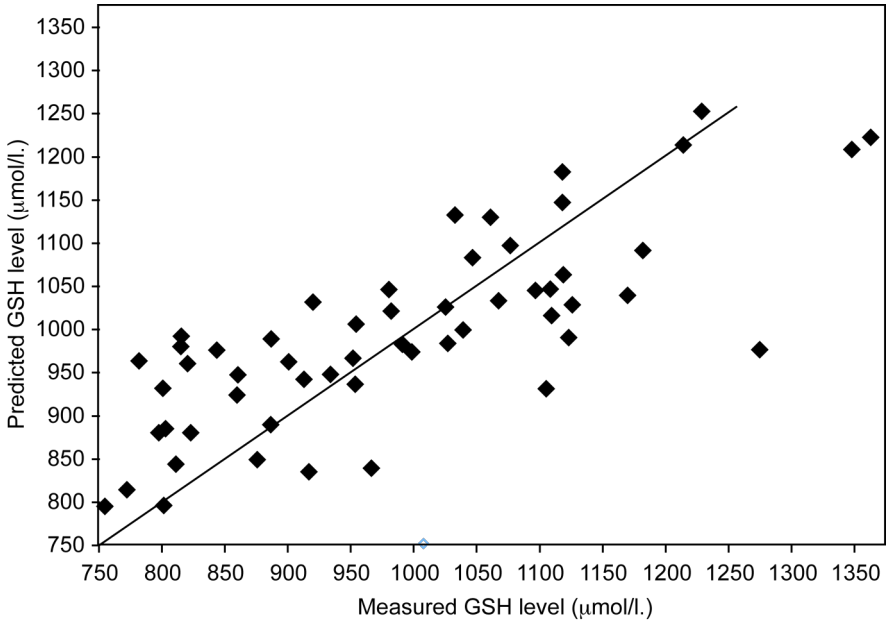


Figure 7
Prediction of GSH concentration from 3 variables by a model with 4 hidden neurons. Estimated generalization error: 175 μmol/l

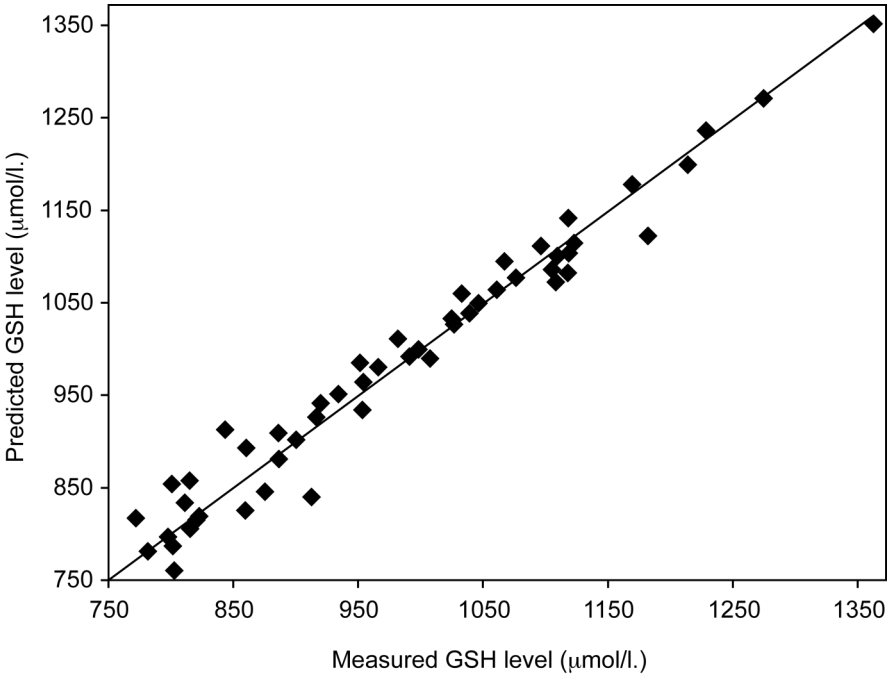


Figure 8
Prediction of GSH concentration from 6 variables by a model with 4 hidden neurons. Estimated generalization error: 153 μmol/l

4.2. Prediction of glutathione and oxidized glutathione from exogenous anti-oxidants

In order to evaluate the relationship between vitamins, oligo-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 represent the probability of the selected variable being more relevant than a probe variable. Figure 9 and Figure 10 show that both quantities can be predicted with satisfactory accuracy. The generalization errors, estimated by virtual leave-one-out, are 177 µmol/l and 1.01 log unit. respectively.

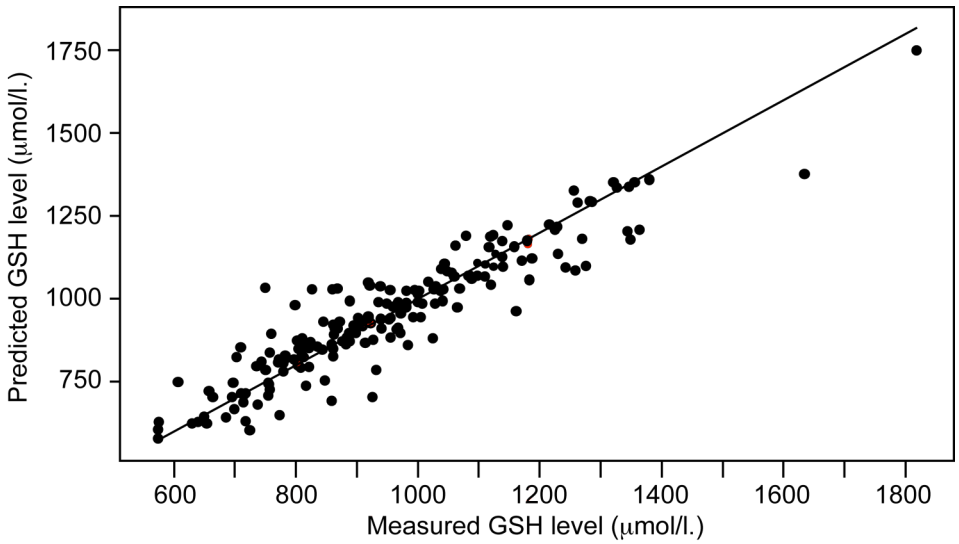


Figure 9
Prediction of glutathione concentration from exogenous antioxidants (199 examples).

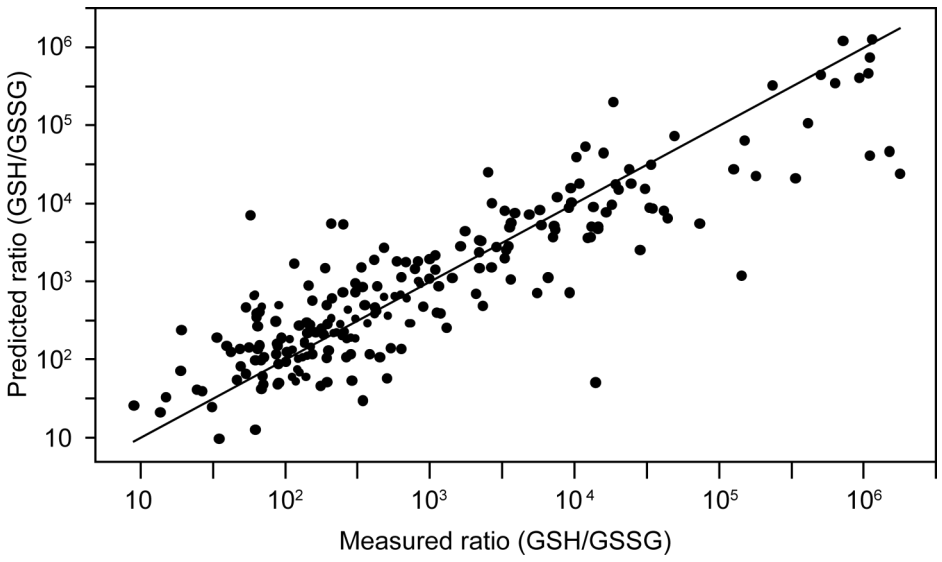


Figure 10
Prediction of the ratio of the concentration of glutathione to the concentration of oxidized glutathione, from exogenous antioxidants (215 examples)

4.3. Prediction of strong markers of oxidative stress: ratio 8-OH-dG/creatinine and oxidized LDL

In the present section, we show that the proposed approach allows the prediction of two strong markers 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 Figure 11 and Figure 12. For the (8-OH-dG/creatinine) concentration ratio, the selected variables 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 5.3. 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%), vitamin C (98%), GSSG (94%), selenium (89%), GSH (81%), and Cu/Zn concentration ratio. The estimated generalization error was 0.6 log units.

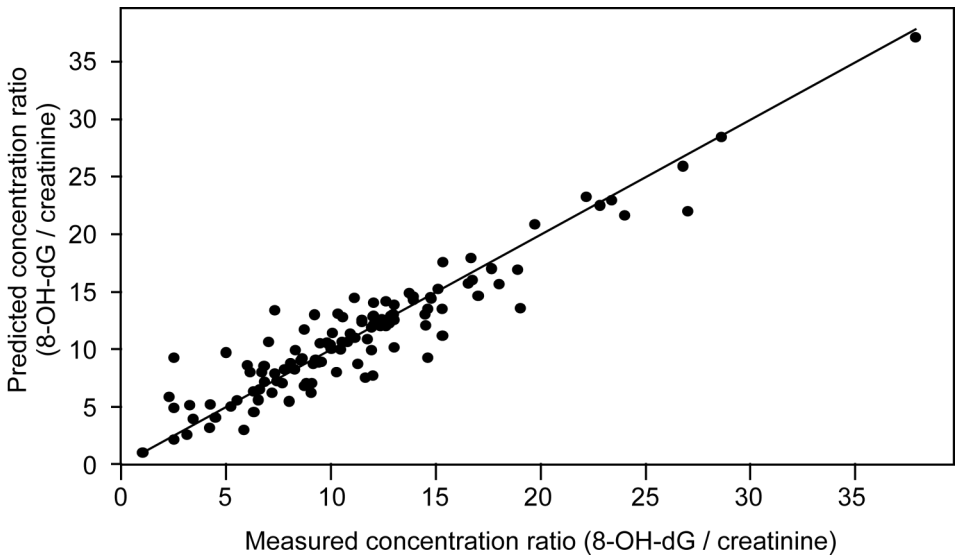


Figure 11
Prediction of the concentration ratio of 8-OH-dG to creatinine (121 examples).

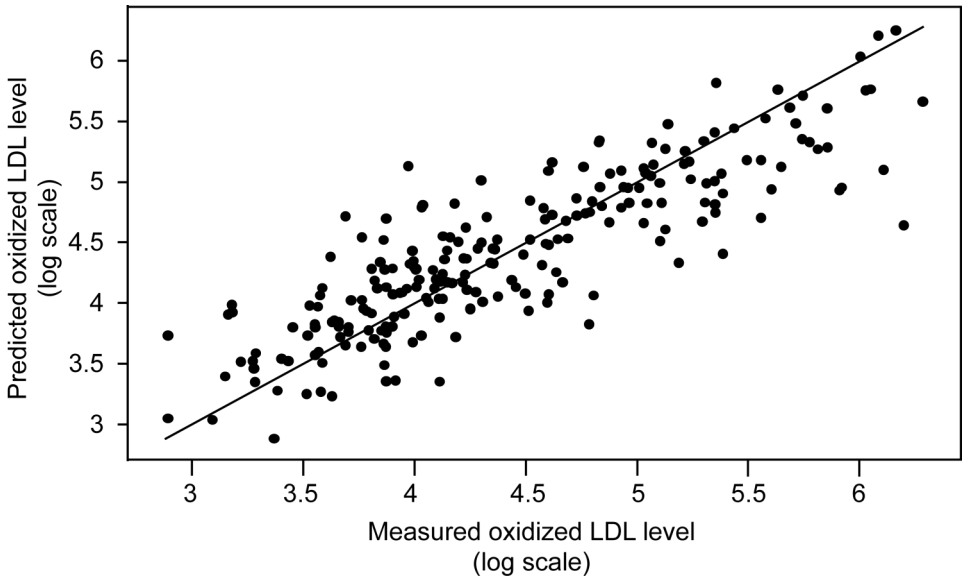


Figure 12
Prediction of the concentration of oxidized LDL (230 examples).

5. Discussion

For the first time, the present study has validated the feasibility of predicting concentrations of markers of oxidative stress, from 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. It addresses the question of the choice of pertinent oxidative stress markers in the context of chronic inflammatory disease, and highlights three clusters of biomarkers: exogenous markers of antioxidant status (vitamins E and C, copper, zinc, selenium, thiols), endogenous markers (GSH and GSSG), and terminal markers of oxidative damage (oxidized LDL and 8-OHdG). The innovative application of a machine learning approach to the prediction of oxidative stress allows us for the first time to predict abnormalities in markers of one group relative to marker abnormalities grouping another.

The critical role of methodological tools including model selection, variable selection, and confidence interval estimation has been demonstrated. From the machine learning point of view, the main open question is the following: are the present results optimal or can they be improved, e.g. 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.

Glutathione level can thus be predicted on the basis of levels of vitamins and some oligoelements, and preferentially by selenium, total thiols, copper/zinc ratio, vitamin E concentration, and finally by the vitamin C/vitamin E ratio (Figures 4 and 5). The predictive power of this approach progressively diminishes when the number of markers decreases from 6 to 3 (Figures 7 and 8), thereby implying that a minimal number of associated markers is required for the pertinent prediction of oxidative stress (Figure 9). The GSH/GSSG ratio can equally be predicted under similar conditions, with the exception of the order of predictive power (order of prediction: copper/zinc ratio, selenium, thiols, vitamin E, and finally vitamin C/vitamin E).

The markers 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) can be predicted by others, as exemplified by those required for glutathione prediction; in addition, GSH/GSSG ratio has also a predictive value for biomarkers of lipid and DNA oxidation.

The pertinence level attained leads to a more appropriate choice of oxidative stress markers, and the predictive power allows reduction in the number of markers to be evaluated, thereby resulting in greater technical and economical feasibility. Clearly then, the appropriate choice of markers is essential for an informative and pertinent diagnostic approach. Finally, the choice of biomarkers constitutes a critical feature of clinical studies involving antioxidant supplementation as it provides key information in the efficiency of the therapeutic response.

These informative findings show that it is worth while pursuing this study on a large biobank derived from patient populations of distinct ethnicity, lifestyle and diet. Moreover, the future targeted patient populations must include a wide spectrum of chronic diseases involving inflammation and oxidative stress in order to allow further evaluation of the present innovative approach.

References

[1] B. Halliwell and J.M. Gutteridge, Free Radicals in Biology and Medicine, 3rd edition ed., Clarendon Press, Oxford, 1999.

[2] B.J. Van Lenten, M. Navab, D. Shih, A.M. Fogelman and A.J. Lusis, Trends Cardiovasc Med 11 (2001) 155-161.

[3] R. Stocker and J.F. Keaney, Jr., Physiol. Rev. 84 (2004) 1381-1478.

[4] K.J. Barnham, C.L. Masters and A.I. Bush, Nat Rev Drug Discov 3 (2004) 205-14.

[5] R.A. Floyd and K. Hensley, Neurobiol Aging 23 (2002) 795-807.

[6] R. Clarke and J. Armitage, Cardiovasc Drugs Ther. 16 (2002) 411-5.

[7] I.D. Coulter, M.L. Hardy, S.C. Morton, L.G. Hilton, W. Tu, D. Valentine and P.G. Shekelle, J Gen Intern Med. 21 (2006) 735-44.

[8] D.Q. Pham and R. Plakogiannis, Ann Pharmacother. 39 (2005) 1870-8. Epub 2005 Sep 27.

[9] S.A. Stanner, J. Hughes, C.N. Kelly and J. Buttriss, Public Health Nutr. 7 (2004) 407-22.

[10] A. Dutta and S.K. Dutta, J Am Coll Nutr. 22 (2003) 258-68.
[11] S.J. Padayatty, A. Katz, Y. Wang, P. Eck, O. Kwon, J.H. Lee, S. Chen, C. Corpe, A. Dutta, S.K. Dutta and M. Levine, J Am Coll Nutr. 22 (2003) 18-35.
[12] G. Bjelakovic, D. Nikolova, L.L. Gluud, R.G. Simonetti and C. Gluud, Jama 297 (2007) 842-57.
[13] J.L. Witztum and D. Steinberg, Trends Cardiovasc Med 11 (2001) 93-102.
[14] J.W. Heinecke, Arterioscler Thromb Vasc Biol 21 (2001) 1261-4.
[15] S. Hercberg, P. Galan, P. Preziosi, S. Bertrais, L. Mennen, D. Malvy, A.M. Roussel, A. Favier and S. Briancon, Arch Intern Med 164 (2004) 2335-42.
[16] P. Therond, D. Bonnefont-Rousselot, A. Davit-Spraul, M. Conti and A. Legrand, Curr Opin Clin Nutr Metab Care 3 (2000) 373-84.
[17] I. Dalle-Donne, R. Rossi, R. Colombo, D. Giustarini and A. Milzani, Clin Chem 52 (2006) 601-623.
[18] M.B. Kadiiska, B.C. Gladen, D.D. Baird, D. Germolec, L.B. Graham, C.E. Parker, A. Nyska, J.T. Wachsman, B.N. Ames, S. Basu, N. Brot, G.A. Fitzgerald, R.A. Floyd, M. George, J.W. Heinecke, G.E. Hatch, K. Hensley, J.A. Lawson, L.J. Marnett, J.D. Morrow, D.M. Murray, J. Plastaras, L.J. Roberts, 2nd, J. Rokach, M.K. Shigenaga, R.S. Sohal, J. Sun, R.R. Tice, D.H. Van Thiel, D. Wellner, P.B. Walter, K.B. Tomer, R.P. Mason and J.C. Barrett, Free Radic Biol Med 38 (2005) 698-710.
[19] M.B. Kadiiska, B.C. Gladen, D.D. Baird, L.B. Graham, C.E. Parker, B.N. Ames, S. Basu, G.A. Fitzgerald, J.A. Lawson, L.J. Marnett, J.D. Morrow, D.M. Murray, J. Plastaras, L.J. Roberts, 2nd, J. Rokach, M.K. Shigenaga, J. Sun, P.B. Walter, K.B. Tomer, J.C. Barrett and R.P. Mason, Free Radic Biol Med 38 (2005) 711-8.
[20] S.T. Omaye, J.D. Turnbull and H.E. Sauberlich, Methods Enzymol 62 (1979) 3-11.
[21] B. Zhao, S.Y. Tham, J. Lu, M.H. Lai, L.K. Lee and S.M. Moochhala, J Pharm Pharm Sci 7 (2004) 200-4.
[22] S. Sturup, R.B. Hayes and U. Peters, Anal Bioanal Chem 381 (2005) 686-94.
[23] F. Tietze, Anal Biochem 27 (1969) 502-22.
[24] G.L. Ellman, Arch Biochem Biophys 82 (1959) 70-7.
[25] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer Verlag; 1999.
[26] Bishop, C. M. *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press; 1995
[27] Dreyfus, G. *Neural Networks, Methodology and Applications.* Springer Verlag; 2005.
[28] Stoppiglia, H.; Dreyfus G.; Dubois R.; Oussar Y. Ranking a Random Feature for Variable and Feature Selection. *Journal of Machine Learning Research* **3:**1399-1414; 2003.
[29] Chen, S.; Billings, S. A.; Luo, W. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control* **50:** 1873-1896; 1989.
[30] Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L.A., eds. *Feature Extraction: Foundations and Applications,* Studies in Fuzziness and Soft Computing Series **207**. Springer Verlag; 2006.
[31] Monari, G., Dreyfus, G. Local Overfitting Control via Leverages. *Neural Computation* **14:** 1481-1506; 2002.
[32] Bates, D. B., Watts, D. G. *Nonlinear Regression Analysis and its Applications*, Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons; 1988.