**OPTICAL IDENTIFICATION OF SUBJECTS AT HIGH RISK FOR DEVELOPING BREAST CANCER**

Paola Taroni¹, Giovanna Quarto¹, Antonio Pifferi¹, Francesca Ieva², Anna Maria Paganoni², Francesca Abbate³, Nicola Balestrieri⁴, Simona Menna³, Enrico Cassano³ and Rinaldo Cubeddu¹

¹ Dipartimento di Fisica
Politecnico di Milano
P.za Leonardo da Vinci 32, 20133 Milano, Italy
(e-mail: paola.taroni@polimi.it)

² MOX - Dipartimento di Matematica
Politecnico di Milano
P.za Leonardo da Vinci 32, 20133 Milano, Italy

³ European Institute of Oncology
Breast Imaging Unit
Via G. Ripamonti, 435, 20141 Milano, Italy

⁴ European Institute of Oncology
Department of Radiology
Via G. Ripamonti, 435, 20141 Milano, Italy

**ABSTRACT.** Breast density is a strong independent risk factor for developing breast cancer. Diffuse optics can estimate tissue composition and derive information on microscopic tissue structure. Taking advantage of this potential, subjects at high risk for developing breast cancer for their high breast density can be effectively identified fitting a logistic regression model to time-resolved multi-wavelength optical data.

**1 INTRODUCTION**

Breast cancer is a leading cause of death in women and a major health burden worldwide (Bray et al. (2004)). Breast density is a recognized strong and independent risk factor for developing breast cancer (McCormack et al. (2006)). It is routinely assessed based on the radiological appearance of breast tissue (Cummings et al. (2009)), thus relying on the use of ionizing radiation. A means for its non-invasive estimate could be broadly and safely applied, even to young women that generally do not undergo X-ray mammography, leading to more effective design of personalized screening paths and prevention strategies for subjects at high risk. Optical techniques are sensitive to tissue composition and structure, and extensive clinical trials showed that optical data, obtained performing continuous wave diffuse optical spectroscopy and interpreted using principal component analysis, strongly correlate with quantitative mammographic features (Blackmore et al. (2008)). We have also obtained encouraging preliminary results (Taroni et al. (2012)) with our portable instrument for time domain optical mammography performed at 7 wavelengths between 635 and 1060 nm (Taroni et al. (2009)). In the present work, we further analyze optically derived parameters using a logistic regression model to devise an optimal way of identifying high-risk women.
2 Experimental set-up and procedures

2.1 Patient study

The Institutional Review Board at the European Institute of Oncology approved the clinical study. Written informed consent is obtained from all the participants. The study has twofold aim: i) the non-invasive assessment of breast density by optical means, and ii) the optical characterization of malignant and benign lesions. Up to now, 199 subjects enrolled. An experienced radiologist evaluated the mammographic density using BI-RADS categories (I - almost entirely fatty, II - scattered fibroglandular densities, III - heterogeneously dense, IV - extremely dense) for 147 patients (age = 52 ± 12 y, BMI = 23.6 ± 3.9 kg).

2.2 Experimental set-up and data analysis

The instrument is designed to collect projection images in compressed breast geometry. Time-resolved transmittance measurements are performed at seven wavelengths (635, 685, 785, 905, 930, 975, 1060 nm) using picosecond pulsed diode lasers and two computer boards for time-correlated single photon counting. The compressed breast is raster-scanned continuously, moving the illumination fiber and collecting bundle in tandem and recording data every millimeter. Optical images are routinely acquired from both breasts in cranio-caudal and medio-lateral oblique (45°) views. Absorption and reduced scattering coefficients at each wavelength are estimated by fitting the experimental data to an analytical solution of the diffusion approximation (with the extrapolated boundary condition) for an infinite homogeneous slab. Information on tissue composition and structure is obtained directly from time-resolved transmittance curves measured at 7 wavelengths. The Beer law is used to relate the absorption properties to the concentrations of the main tissue constituents. The scattering properties are modeled through the simple approximation to Mie theory: $\mu_s^r(\lambda) = a(\lambda/\lambda_0)^b$, where $\lambda_0 = 600$ nm and the scattering amplitude $a$ is the reduced scattering coefficient $\mu_s^r(\lambda_0)$. A spectrally constrained global fitting procedure is applied, where free parameters are the concentrations of oxy- and deoxy-hemoglobin, water, lipids, and collagen, together with the scattering amplitude $a$ and power $b$. For the present study, tissue composition and scattering parameters were averaged over each image, excluding regions close to the boundary of the compressed breast, which cannot correctly be described by the infinite slab model. We performed a descriptive analysis of data, as well as non-parametric Wilcoxon tests to compare the optical measurements distributions between subjects in BI-RADS categories 1 to 3 and subjects in BI-RADS category 4. Then the best regression logistic model for the risk probability has been chosen via a stepwise variables selection minimizing the AIC (Akaike Information Criterion). This model has been used to classify patients to high or low risk categories.
The present study aims at identifying a simple and automatic way to effectively discriminate women that are at high risk for developing breast cancer due to high breast density, based on information that can be derived from non-invasive optical measurements. Linear correlation between optically derived parameters was initially investigated. The strongest (negative) correlation is observed between lipid and water content, but negative correlation is also evident between lipid and collagen content. Both observations are in agreement with what expected based on breast tissue composition: moving from adipose to fibroglandular breasts involves a progressive decrease in adipose tissue, with high lipid content, which is replaced by connective and epithelial tissue, richer in water and collagen. Marked correlation also exists between the scattering amplitude $a$ and the concentrations of all major tissue constituents, not only water and lipid, but also collagen, as expected since $a$ is directly related to the number of effective scattering centers, which increases upon increasing constituent concentrations.

To develop a procedure for the identification of high-risk women, mammographic density was dichotomized, comparing subjects in BI-RADS categories 1 to 3 to subjects in BI-RADS category 4, the latter being at significantly higher risk than all the others and thus being considered as representative of the high-risk population (McCormack et al. (2006)). The p-values of the Wilcoxon test showed that total hemoglobin content, lipid, water, collagen, $a$ and $b$ are significantly different in the two populations considered (at least p < 0.001), while the oxygen saturation level is not. Best fit of a logistic regression model led to the following:

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 \text{Collagen}_i + \alpha_2 a_i + \alpha_3 b_i,$$

where $p_i$ is the probability of belonging to BI-RADS category 4 (high-risk). Table 1 shows the output of the fitted regression logistic model (point estimates of the coefficients, related standard errors, z-statistics and p-values of testing their significance in the model). The Brier’s score, i.e. the mean square difference between outcome and estimated probability, is equal to 0.095. Based on Eq. 1 and Table 1, the probability of belonging to the high-risk category depends on collagen concentration and on both scattering parameters. In particular, the strongest dependence occurs for the scattering slope $b$. Actually, we have recently observed that the scattering slope measured in vivo depends strongly on collagen content and possibly structure (Bray et al. (2004)). Collagen content also shows a significant linear correlation with the scattering amplitude $a$, as mentioned above. Thus, both directly and indirectly, collagen seems to be the most crucial feature for the identification of subjects with high breast density.
density. We use model 1 to classify patients: if the estimated probability is greater than 0.5 we classify the subject as a high-risk patient. The results are reported in Table 2, where "true" refers to risk classification based on mammographic assessment (BI-RADS categories) and "classified" to risk as predicted based logistic regression fitted on optical data.

<table>
<thead>
<tr>
<th></th>
<th>Classified as Low Risk (Optical)</th>
<th>Classified as High Risk (Optical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Low Risk</td>
<td>105</td>
<td>7</td>
</tr>
<tr>
<td>True High Risk</td>
<td>11</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 2. Misclassification matrix.

The total misclassification error is 12.3%, corresponding to a simple kappa of 0.84. The latter should be compared with the reproducibility of BI-RADS measures among radiologists and even intra-radiologist. Specifically, the intra-rater agreement is of 77%, leading to a simple kappa of 0.58 (Spayne et al. (2012). Thus, the optical classification seems to perform favorably.

4 CONCLUSION

A logistic regression model was fitted to optically derived tissue parameters with the aim of identifying women at high risk for developing breast cancer because of their high breast density. Promising preliminary results were obtained in the optical classification of high-risk subjects, and collagen proved to be the key parameter for the classification, either directly (collagen content) or indirectly (through scattering amplitude and slope). The relevance of collagen is in agreement with what expected based on breast anatomy and physiology, and opens up the possibility of a more direct estimate of breast density than presently achieved using X-ray mammography, which is mostly sensitive to water content and not directly to collagen.

REFERENCES


