*STATISTICA, anno LXX, n. 3, 2010*

# SOME ISSUES CONCERNING THE STATISTICAL EVALUATION OF A SCREENING TEST: THE ARFI ULTRASOUND CASE

M. Attanasio, M. Enea, L. Rizzo

## 1. INTRODUCTION

In this paper we deal with some statistical issues concerning the screening tests for the diagnosis of fibrosis in patients with Chronic Hepatitis C (HCV). The prognosis and the clinical management of chronic liver diseases are dependent on the extent of the liver fibrosis. The majority of physicians considers liver biopsy to be the most reliable screening test for HCV (Saleh and Abu-Rashed, 2007). In spite of being invasive, painful, and potentially life-threatening, liver biopsy remains the gold standard for staging liver disease and is usually measured by the METAVIR scoring system. Unfortunately, liver biopsy presents some inconveniences in assessing liver fibrosis such as the inter-observer and inter-procedure variability (due to not complete consistency in defining pathological features, different technical processing of the specimens, etc. Shiha *et al.*, 2009) and the limited volume of the liver tested by the biopsy (the bioptical sample is just 1/50000 of the entire liver (Bravo *et al.*, 2001). Liver biopsy, because of its limitations and risks, is no longer considered the first-line indicator of liver injury, and consequently there has been an intensive search for alternative non-invasive methods for staging of the disease with many markers having been developed as non-invasive alternatives (Carey and Carey, 2010). In this paper we explore some of these alternative techniques and consider results from methodologies derived from current liver ultrasound techniques as Transient Elastography (TE) (Sandrin *et al.*, 2003) and Acoustic Radiation Force Impulse (ARFI) (Palmeri *et al.*, 2008). TE is a rapid and user-friendly device that can be easily used at the patient bedside or in an outpatient clinic with immediate results and good reproducibility (Talwalkar *et al.*, 2007, Friedrich-Rust *et al.*, 2008), even if the technique is burdened by a series of confounding factors, which might reduce its diagnostic accuracy (Arena *et al.*, 2008). Recently, ARFI has been also used for the diagnosis of liver fibrosis (Castera, 2009). The general objective of this paper is to provide statistical tools to measure the performance of diagnostic devices, comparing results from 130 patients who have experienced both biopsy and ultrasound examination (TE and ARFI). Our study focuses on the assessment of the ARFI performance in diagnosing liver fibrosis. We have first assessed the performance of the ARFI in comparison to other screening tests and have then provided measures of reliability of the ARFI diagnosis compared to the ones offered by the biopsy. Our analysis does not address the ROC curve analysis because our aim does not include the cut-off determination (Sullivan Pepe, 2003). In Section 2, we describe the aim, the data, and the screening tests. In Section 3, we report the comparisons between ARFI and TE versus METAVIR. In Section 4, we briefly describe the bootstrap procedure, the results, and final remarks.

## 2. THE AIM, THE DATA , AND THE SCREENING TESTS

A total of 130 patients from 2 centers (Catania=72 and Palermo=58) were included in the database. They were consecutive patients with suspected HCV: 72 patients were admitted at University Hospital in Catania for the TE and METAVIR and at the outpatient clinic of Catania "Ultrasuoni Rizzo" for the ARFI examination, and, 58 patients were admitted at the Gastroenterology Unit of the University Hospital of Palermo for the three examinations. Therefore data can be considered as coming from an observational study. The aim of the paper is twofold:

1. to compare the two ultrasound competitors (TE and ARFI) versus the biopsy (METAVIR) in terms of agreement;
2. to construct a table which provides the probability, within some fixed error intervals, of getting the "correct" diagnosis.This table may be useful for the physician because in this way he/she is able to provide the level of reliability of the screening test. The statistical method used to calculate the reliability index (probability) is the resampling method bootstrap.

In order to achieve the first aim of our paper, we will compare results from 130 patients, as already said, who had had both the biopsy and ultrasound machines examinations, TE and ARFI. For each patient three measures are available:

- METAVIR is an ordinal five-point scale (F0-F1, F2, F3 and F4). The first two categories F0-F1 are aggregated for simplicity and correspond to a patient which usually will not be treated. The categories F2 and F3 correspond to intermediate severity of the fibrosis and in these cases physicians advice treatment. The last category, F4, corresponds to the most severe stage of fibrosis, that is, the patient is cirrhotic.
- TE provides a quantitative measure in kPa, usually ranging from 0 to 50. We adopt the usual conversion scale into the METAVIR scoring system (Sandrin *et al.*, 2003):

| | F0-F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| TE(kPa) | <7.0 | [7.0, 8.8) | [8.8, 12.0) | ≥12.0 |

- ARFI provides a quantitative measure in m/s. Assuming, by clinical practice, that values below 1.3 correspond to untreatable patients and values equal or greater than 2.0 correspond to cirrhotic patients, we adopt a conversion scale just dividing with equally spaced intervals the interval between 1.3 and 2.0. This conversion corresponds to the well known quadratic conversion between kPa and m/s.

| | F0-F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| ARFI(m/s) | <1.3 | [1.3, 1.7) | [1.7, 2.0) | ≥2.0 |

In order to achieve the second aim of our paper we need to develop a sampling scheme to collect data. The rationale stands in the physiopathology of the liver: our assumption is that the disease (the liver fibrosis, which is measured by the stiffness of the liver tissue) seems to be related to the liver heterogeneity, that is the stiffness is present "haphazardly" in the human liver. Because of this assumption, the easiest diagnoses are obtained when there is homogeneity: either when the liver is not stiff at all or when it is entirely stiff.

In these two extreme cases, one measurement alone may be sufficient to get enough information for a diagnosis. The problem and the interest lie in analyzing the intermediate cases because stiffness spreads out evenly in the liver and a measure of its variability may be useful to get better diagnoses. From this point of view, ARFI presents a major advantage comparing to the other two methods because it records multiple observations for each patient, while METAVIR and TE report just a single value representing the central tendency (even if the TE may be set in order to provide several measurements of the same spot). Assuming that liver heterogeneity is present in diseased patients, we need to settle a sampling scheme in advance in order to guarantee a "good representation" of the liver. This sampling scheme will be described just before considering the bootstrap procedure. Table 1 reports the cross-classification of the patients according to the three screening tests.

TABLE 1

*Three-way contingency table of patients according Metavir (i), Arfi (j) and Te (k)*

| Metavir | Arfi | Te | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| 1 | 1 | 25 | 4 | 1 | 0 |
| | 2 | 6 | 6 | 0 | 2 |
| | 3 | 0 | 1 | 0 | 0 |
| | 4 | 0 | 0 | 0 | 0 |
| 2 | 1 | 9 | 3 | 2 | 0 |
| | 2 | 6 | 1 | 1 | 1 |
| | 3 | 3 | 1 | 2 | 2 |
| | 4 | 0 | 1 | 1 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 |
| | 2 | 1 | 1 | 0 | 1 |
| | 3 | 1 | 4 | 1 | 1 |
| | 4 | 1 | 0 | 1 | 11 |
| 4 | 1 | 0 | 0 | 1 | 0 |
| | 2 | 0 | 0 | 0 | 1 |
| | 3 | 0 | 0 | 1 | 2 |
| | 4 | 4 | 0 | 3 | 17 |

## 3. TE AND ARFI: TWO COMPETITORS VERSUS THE GOLD STANDARD METAVIR

The statistical evaluation of the diagnostic accuracy among several screening tests is conducted in an unusual way: we have two competitors (TE and ARFI) versus a gold standard (METAVIR). Both TE and ARFI are on the same ground, because their costs and typology are comparable, while METAVIR is less preferable for reasons previously stated even though is still considered the gold standard in the diagnosis of liver disease. Firstly, in order to compare these three devices, we calculate the Weighted Kappa (Cohen, 1968)

$$K = 1 - \frac{\sum_{i=1}^{k}\sum_{i=1}^{k} w_{ij} p_{ij}}{\sum_{i=1}^{k}\sum_{i=1}^{k} w_{ij} \pi_{ij}}$$

of TE versus METAVIR and of ARFI versus METAVIR, where $w_{ij}$ are weights for the observed entries $p_{ij}$ or the expected values $\pi_{ij}$. We use the R function cohen.kappa (package psych), which provides zero weights as default for the diagonal elements and weights equal to the squared distances for the off diagonal ones, that is $w_{ij} = (i-j)^2/(k-1)^2, i, j = 1,...,k.$ The calculations are reported in Tables 2 and 3.

TABLE 2

*Cohen Kappa and Weighted Kappa correlation coeffcients and confidence boundaries of METAVIR vs TE*

| | lower | estimate | upper |
|---|---|---|---|
| weighted kappa | 0.14 | 0.25 | 0.35 |
| unweighted kappa | 0.52 | 0.64 | 0.76 |

TABLE 3

*Cohen Kappa and Weighted Kappa correlation coeffcients and confidence boundaries of METAVIR vs ARFI*

| | lower | estimate | upper |
|---|---|---|---|
| weighted kappa | 0.26 | 0.37 | 0.48 |
| unweighted kappa | 0.72 | 0.79 | 0.86 |

The results of the tests show how ARFI performs better than TE, with respect to METAVIR. But Cohen's Kappa presents several inconveniences: its interpretation is not straightforward (the value itself defines just large categories of agreement and the weighting system is rather arbitrary), it does not provide specific values for the stages of the disease. In order to answer to some of these inconveniences, we suggest another measure of agreement $OR(Ag)$, based on the odds ratio rationale. It is an asymmetrical test appropriate for our case, which simultaneously compares the overall agreement between two competitors conditioned to the levels of the gold standard METAVIR. In order to calculate the $OR(Ag)$'s we need to build another table (Table 4) obtained splitting Table 1.

TABLE 4

*Concordance/Discordance tables of* TE *vs* ARFI *conditioned to* METAVIR

| Metavir=1 | | Te | | |
|---|---|---|---|---|
| | | 1 | ≠1 | Total |
| Arfi | 1 | 25 | 5 | 30 |
| | ≠1 | 6 | 9 | 15 |
| | Total | 31 | 14 | 45 |

| Metavir=2 | | Te | | |
|---|---|---|---|---|
| | | 2 | ≠2 | Total |
| Arfi | 2 | 1 | 8 | 9 |
| | ≠2 | 5 | 19 | 24 |
| | Total | 6 | 27 | 33 |

| Metavir=3 | | Te | | |
|---|---|---|---|---|
| | | 3 | ≠3 | Total |
| Arfi | 3 | 1 | 6 | 7 |
| | ≠3 | 1 | 15 | 16 |
| | Total | 2 | 21 | 23 |

| Metavir=4 | | Te | | |
|---|---|---|---|---|
| | | 4 | ≠4 | Total |
| Arfi | 4 | 17 | 7 | 24 |
| | ≠4 | 3 | 2 | 5 |
| | Total | 20 | 9 | 29 |

The elements of each sub-table are:

1. (1, 1) is the number of concordant patients in the three examinations;
2. (2, 2) is the number of patients for which both ARFI and TE are discordant with METAVIR;
3. (1, 2) and (2, 1) are the number of patients for which, respectively, METAVIR is concordant with ARFI (but discordant with TE), and METAVIR is concordant with TE (but discordant with ARFI).

We consider the margin-based odds $\Omega_{i1+} = n_{i1+} / n_{i2+}$ I and $\Omega_{i+1} = n_{i+1} / n_{i+2}$, $i$ = 1,...4, with the subscripts in the following order: METAVIR, ARFI, and TE. The ratio $OR_i(Ag) = \Omega_{i1+} / \Omega_{i+1}$ compares the two screening tests and each $OR_i(Ag)$ ($i$ = 1,...4) and provides a measure of agreement conditioned to the METAVIR level. The $OR_i(Ag)$'s have the usual interpretation: a value greater than 1 means that ARFI has a better performance than TE in the i-th category. However, in order to get a better asymptotic approximation of the $OR_i(Ag)$' s distribution, we consider the log-odds ratio:

$$\log OR_i(Ag) = \log \Omega_{i1+} - \log \Omega_{+i1}.$$

The asymptotic variance for the i-th category of METAVIR are estimated as in Hwang and Biswas (2008):

$$\operatorname{var}(\log OR_i(Ag)) = \operatorname{var}(\log \Omega_{i1+}) + \operatorname{var}(\log \Omega_{+i1}) - 2\operatorname{cov}(\log \Omega_{i1+}, \log \Omega_{+i1})$$

$$\approx \frac{1}{n_{i1+}} + \frac{1}{n_{i2+}} + \frac{1}{n_{i+1}} + \frac{1}{n_{i+2}} - 2\rho_i \sqrt{\left(\frac{1}{n_{i1+}} + \frac{1}{n_{i2+}}\right)\left(\frac{1}{n_{i+1}} + \frac{1}{n_{i+2}}\right)},$$

where $\rho_i$ is the correlation coefficient between ARFI and TE inside the *i*-th category. An estimate of $\rho_i$ is given by

$$\hat{\rho}_i = \hat{\theta}_i \sqrt{\frac{p_{i1+}p_{i2+}}{p_{i+1}p_{i+2}}}$$

with $\hat{\theta}_i = p_{i12} / p_{i1+} - p_{i21} / p_{i+1}$.

The final index is just a weighted mean of all the gains, that is, an overall concordance measure in which the weights of the log odds ratios are given by the sample proportions:

$$\log OR(Ag) = \sum_{i=1}^{k} \log\left(\frac{\Omega_{i1+}}{\Omega_{i+1}}\right) p_{i++},$$

$$\log OR(Ag) = \log(0.90)0.35 + \log(1.69)0.25 + \log(4.59)0.18 + \log(2.16)0.22 = 0.54,$$

$$\text{var}(\log OR(Ag)) \approx \sum_{i=1}^{k} \text{var}\left(\log \frac{\Omega_{i1+}}{\Omega_{i+1}}\right) p_{i++}^2 = 0{,}078.$$

The interpretation of the index is straightforward because it has the intuitive meaning of the odds ratio. An alternative to $OR(Ag)$ may be given by McNemar test (1947), which uses matched pairs of objects difference of proportions of two dependent proportions. In a similar way, it is possible to get an average in order to obtain an overall measure. The indexes with their standard errors and p-values are reported in Table 5.

TABLE 5

*Agreement assessment: odds, odds ratios, log-odds ratios, standard errors and p-values*

| METAVIR | $n_i$ | $\Omega_{i1+}$ | $\Omega_{i+1}$ | $\Omega_{i1+} / \Omega_{i+1}$ | $\log(OR_i(Ag))$ | s.e. | p-value |
|---|---|---|---|---|---|---|---|
| 1 | 45 | 30/15 | 31/14 | 0.90 | -0.10 | 0.337 | 0.763 |
| 2 | 33 | 9/24 | 6/27 | 1.69 | 0.52 | 0.629 | 0.406 |
| 3 | 23 | 7/16 | 2/21 | 4.59 | 1.52 | 0.815 | 0.061 |
| 4 | 29 | 24/5 | 20/9 | 2.16 | 0.77 | 0.606 | 0.204 |
| Overall | 130 | - | - | - | 0.54 | 0.280 | 0.054 |

The results report the “gain” given by ARFI compared to TE for each stage of stiffness, but in the first row. Therefore ARFI is preferable to TE in almost all the categories, even if its *p*-values are not close to 0.05, but for METAVIR equal to 3. However, we should have to consider that the sample size is rather small in each category. Interestingly, $\log OR(Ag)$ = 0.54, representing the overall “gain”of ARFI compared to TE, has the smallest standard error.

A limitation of the proposed measure is that it does not take into account of the ordering of the data. Thus, this measure should be viewed as an alternative to the usual Cohen's Kappa rather than to its weighted version.

## 4. THE SAMPLING SCHEME

The determination of the "optimal" number of measurements presents the usual trade-off between reliability and cost. It is obvious that it is preferable an ARFI examination with a limited number of measurements. Therefore it is crucial to find out the sample size and its corresponding significance level to provide an useful information for the physician. In order to get such information, we need a sampling scheme which considers some important features of the liver as well as the device. Therefore, the sampling scheme has to follow a random scheme which takes into account the liver features and the ARFI potentials. To apply the resampling method bootstrap, we need to construct a random sequence of observations for each patient, that is, the series of measurements have not to be influenced by the observer. To get those measurements, we suggest how to move the ARFI transducer, assuming that the liver fibrosis may vary according to the severity of the disease. The question is: which are the "best" parts of the liver to be sampled? The sampling scheme type is random stratified, in which the stratification variables are suggested by the experts. They are:

| | |
|---|---|
| 1. *Anatomic Region (AR*) | 3 categories; |
| 2. *Intercostal Space (IS*) | 3 categories; |
| 3. *Breath (B)* | 2 categories; |
| 4. *Depth (D*) | 3 categories. |

| | CATEGORIES | | |
|---|---|---|---|
| AR | The space between emiclavear and anterior ascilla ligne, in supine decubitus | The space between anterior axillary line and midaxillary line, in left decubitus | The space between midaxillary line and posterior axillary line, in prone decubitus |
| IS | 8 | 9 | 10 |
| B | Inspiration | Expiration | |
| D | 3 cm from body surface | 4 cm from body surface | 5 cm from body surface |

For each patient we get $H$ repeated measurements in different spots of the liver. Each point $X$ of the liver is pointed by the vector $x(ar, is, b, d)$, randomly drawn. In this way it is necessary that the observer constructs in advance an array (4, *H*) for each patient. The final array of the measurements is:

| | *1-st measurement* | *2-nd measurement* | ... | *H-th measurement* |
|---|---|---|---|---|
| *Patient 1* | $x(ar_{1,1}, is_{1,1}, b_{1,1}, d_{1,1})$ | $x(ar_{1,2}, is_{1,2}, b_{1,2}, d_{1,2})$ | ... | $x(ar_{1,H}, is_{1,H}, b_{1,H}, d_{1,H})$ |
| *Patient 2* | $x(ar_{2,1}, is_{2,1}, b_{2,1}, d_{2,1})$ | $x(ar_{2,2}, is_{2,2}, b_{2,2}, d_{2,2})$ | ... | $x(ar_{1,H}, is_{1,H}, b_{1,H}, d_{1,H})$ |
| ... | | | | |
| *Patient K* | $x(ar_{K,1}, is_{K,1}, b_{K,1}, d_{K,1})$ | $x(ar_{K,2}, is_{K,2}, b_{K,2}, d_{K,2})$ | ... | $x(ar_{K,H}, is_{K,H}, b_{K,H}, d_{K,H})$ |

In practice, the minimum number of measurements *n* is approximately 15 and it usually increases as variability increases. More than 70 measurements are not feasible.

## 5. DETERMINING THE SAMPLE SIZE *n* FOR THE MEAN

How many ARFI measurements are necessary for an accurate estimate of the population mean for a single patient? That number depends on the variability of the liver stiffness since, as the sample mean increases, the standard deviation increases as well. Figure 1 shows the positive empirical correlation between the sample mean and the sample standard deviation. Therefore, the standard deviation plays an important role in determining the "correct" number of measurements.

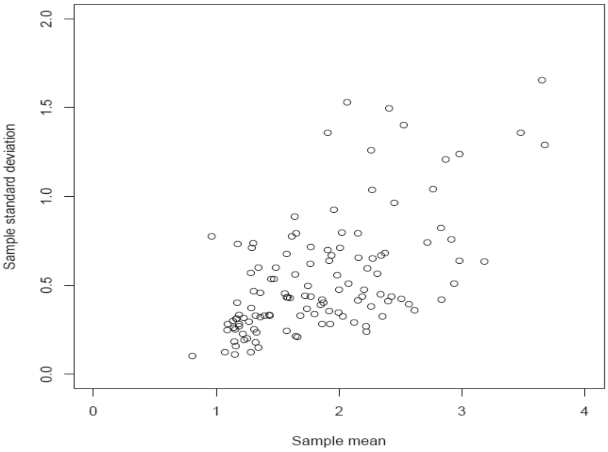

*Figure 1* – Scatterplots of sample mean vs standard deviation for a sample of 130 patients.

In this section, our aim is to make inference for the mean, controlling for different values of *n* and standard deviation. The sample size *n* for the mean depends on the error *r* such that $\Pr(|\overline{X}-\overline{x}|\geq r)=\alpha$, where $\overline{X}$ is the population mean, $\overline{x}$ is the sample mean, and $\alpha$ is the significance level. To determinate the sample size, we can use the formula $n=Z^2\sigma^2/r^2$, where $Z$ is the quantile of the normal distribution at the significance level $\alpha/2$, $\sigma^2$ is the unknown variance of the distribution, and *r* is the absolute error. To estimate $\sigma^2$ we can use the sample variance $S^2$ and to determine the sample size for a relative error (with respect to the sample mean) the usual formula is $n=(Z^2\sigma^2)/(r^2\overline{x}^2)$ (Cochran, 1977). Table 6 reports the theoretical sample sizes for the mean, using the significance levels 0.05 and 0.1 for the normal distribution, controlling for some fixed absolute errors and standard deviations.

TABLE 6

*Theoretical sample sizes for the mean, controlling for absolute errors, standard deviations, and significance levels $\alpha$ =0.05, 0.1*

| | | *r* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ | 1- $\alpha$ | 0.01 | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 | 0.35 | 0.4 |
| 0.1 | 0.95 | 15.4 | 3.8 | 1.7 | 1.0 | 0.6 | 0.4 | 0.3 | 0.2 |
| | 0.90 | 10.8 | 2.7 | 1.2 | 0.7 | 0.4 | 0.3 | 0.2 | 0.2 |
| 0.2 | 0.95 | 61.5 | 15.4 | 6.8 | 3.8 | 2.5 | 1.7 | 1.3 | 1.0 |
| | 0.90 | 43.3 | 10.8 | 4.8 | 2.7 | 1.7 | 1.2 | 0.9 | 0.7 |
| 0.5 | 0.95 | 384.2 | 96.0 | 42.4 | 24.0 | 15.4 | 10.7 | 7.8 | 6.0 |
| | 0.90 | 270.5 | 67.6 | 30.1 | 16.9 | 10.8 | 7.5 | 5.5 | 4.2 |
| 1 | 0.95 | 1536.6 | 384.2 | 170.7 | 96.0 | 61.5 | 42.7 | 31.4 | 24.0 |
| | 0.90 | 1082.2 | 270.5 | 120.2 | 64.6 | 43.3 | 30.1 | 22.1 | 16.9 |
| 1.5 | 0.95 | 3457.4 | 864.4 | 384.2 | 216.1 | 138.3 | 96.0 | 70.6 | 54.0 |
| | 0.90 | 2435.0 | 608.7 | 270.5 | 152.2 | 97.4 | 67.6 | 49.7 | 38.0 |

Although the values reported in Table 6 may be obtained theoretically, in presence of non-normal distributions, the convergence to normality may be slow, and the theoretical quantiles of the normal distribution may be not appropriate. To overcome such inconveniences, the bootstrap method can be used to simulate the distribution of the sample mean, especially in presence of skewed data.

## 5.1 *Bootstrapping the sample mean*

To carry out the simulation, we have chosen five patient profiles, corresponding to increasing sample standard deviations $S$ = 0.01, 0.2, 0.5, 1.0, 1.5. The sampling histograms of the ARFI repeated measurements for these patients are showed in Figure 2.

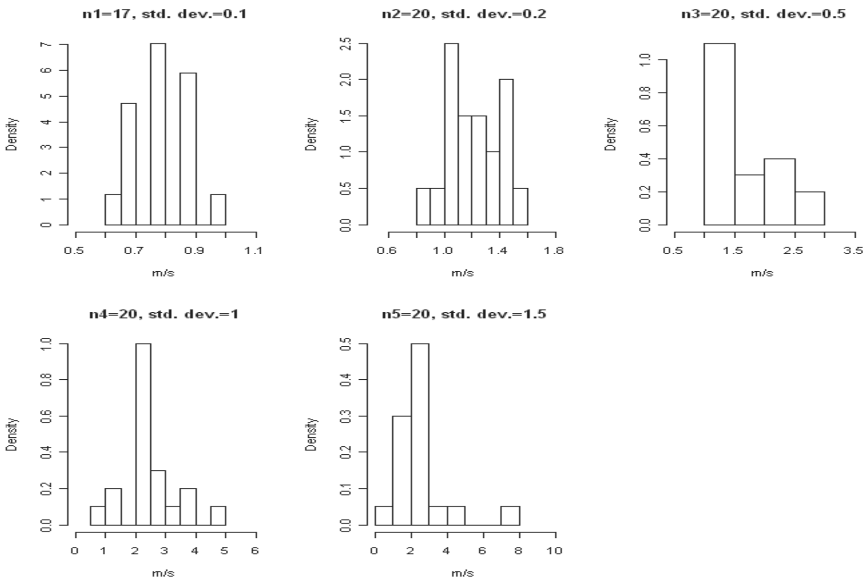

*Figure 2* – Sampling histograms of ARFI repeated measurements for five patients by several sample standard deviations.

For each profile, we simulate the sampling bootstrap distribution for the chosen statistics and calculate the bootstrap variance. Let $\theta^*$ be the bootstrap estimate. An estimate of *r* is just $\tilde{r}_{n^*} = |\theta^*_{n^*} - \bar{x}|$, whereas an estimate of 1-$\alpha$ is the proportion $p^*$ of bootstrap estimates within the interval $\theta^*_{n^*} \pm \tilde{r}_{n^*}$. The corresponding value of $n^*$ for which $p^*$ is closer to the chosen level 1-$\alpha$ is the bootstrap estimate of the sample size *n*. Table 7 reports the coverage proportions $p^*$ for the bootstrap distribution of the sample mean, for fixed error and standard deviation.

TABLE 7

*Coverage proportions $p^*$ for the bootstrap distribution of the sample mean, by fixed error and standard deviation*

| | | $\tilde{r}_{n^*}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| S | 1- $\alpha$ | 0.01 | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 | 0.35 | 0.4 |
| 0.1 | 5 | 0.73 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 10 | 0.88 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 15 | 0.95 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 20 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 25 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.2 | 10 | 0.61 | 0.91 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 20 | 0.74 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 30 | 0.84 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 50 | 0.94 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 60 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 70 | 0.97 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.5 | 10 | 0.23 | 0.46 | 0.66 | 0.80 | 0.88 | 0.95 | 0.97 | 0.99 |
| | 20 | 0.32 | 0.61 | 0.81 | 0.92 | 0.98 | 0.99 | 1.00 | 1.00 |
| | 30 | 0.41 | 0.74 | 0.92 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 40 | 0.48 | 0.81 | 0.95 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 50 | 0.54 | 0.85 | 0.97 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 60 | 0.54 | 0.88 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 70 | 0.61 | 0.91 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1.00 | 10 | 0.13 | 0.25 | 0.38 | 0.49 | 0.60 | 0.68 | 0.75 | 0.82 |
| | 20 | 0.20 | 0.36 | 0.52 | 0.65 | 0.77 | 0.84 | 0.91 | 0.94 |
| | 30 | 0.24 | 0.46 | 0.64 | 0.77 | 0.86 | 0.92 | 0.96 | 0.97 |
| | 40 | 0.24 | 0.47 | 0.68 | 0.81 | 0.91 | 0.96 | 0.98 | 0.99 |
| | 50 | 0.30 | 0.56 | 0.76 | 0.88 | 0.94 | 0.98 | 0.99 | 1.00 |
| | 60 | 0.31 | 0.58 | 0.77 | 0.89 | 0.95 | 0.99 | 1.00 | 1.00 |
| | 70 | 0.33 | 0.63 | 0.82 | 0.93 | 0.97 | 0.99 | 1.00 | 1.00 |
| 1.5 | 10 | 0.08 | 0.15 | 0.22 | 0.30 | 0.39 | 0.48 | 0.57 | 0.63 |
| | 20 | 0.12 | 0.22 | 0.36 | 0.46 | 0.56 | 0.63 | 0.7 | 0.76 |
| | 30 | 0.17 | 0.31 | 0.44 | 0.55 | 0.65 | 0.74 | 0.80 | 0.86 |
| | 40 | 0.16 | 0.33 | 0.48 | 0.60 | 0.71 | 0.79 | 0.88 | 0.92 |
| | 50 | 0.21 | 0.39 | 0.55 | 0.68 | 0.77 | 0.85 | 0.92 | 0.96 |
| | 60 | 0.19 | 0.40 | 0.56 | 0.71 | 0.81 | 0.89 | 0.95 | 0.97 |
| | 70 | 0.24 | 0.44 | 0.62 | 0.75 | 0.86 | 0.93 | 0.98 | 0.99 |

The results are close to the theoretical ones, as we expected. For example, looking at the first block with standard deviation $S$ = 0.1 and $\tilde{r}_{n^*} = 0.05$, 95% of bootstrap means $\theta^*$ are within the interval centered on the sample mean $\bar{x}$, that is within $\bar{x} \pm 0.05$, with a corresponding $n^{*} = 15$, just as in Table 6. In addition, it is possible to get other values of $n^{*}$, not reported in the table, by interpolation. As expected, there are some discrepancies between the theoretical values of n and the bootstrap estimates $n^{*}$. For instance, in the fifth block with $S$ = 1.5, $\tilde{r}_{n^{*}} = 0.35$ and $p^{*} = 0.95$, we get $n^{*} = 60$ whereas the corresponding n is about 71. This bias may be due to at least two reasons: a) the original sample is strongly non-normal (as, for instance, data of the 3-*rd* and the 5-*th* patient); b) the bootstrap sample size is larger than the original sample size. However, for small sample standard deviations the bias of $n^{*}$ is negligible.

## 6. CONCLUSIONS

In this paper we have investigated on the diagnosis performance of a new ultrasound screeening test for the liver fibrosis. The comparison of two competitor screening tests versus a gold standard was dealt with an asymmetrical version of an agreement index based on Odds Ratios for dependent data. This alternative index employs the margins of a matched pairs contingency table. It can be computed both conditioning to the different levels of the disease and averaging such conditioned indexes with weights proportioned to the levels.

Arfi ultrasound performs better than TE but in the first level of the disease, even if these data could be revised by larger sample sizes, especially for the intermediate stages where the diagnosis is more difficult. Further developments of the proposed agreement index for ordinal data may provide better insight. The second scope aims at providing an evaluation grid of the reliability of the measurements according to their size and their standard deviation. In practice, physicians can get a helpful table containing the “correct” number of measurements to be done in order to get a statistical estimate of the true stage (given by the METAVIR), controlling for the mean and the satndard deviation, which are estimayed by the observed measurements. Therefore the sample size *n* for a single patient has been determined for the mean of the measurements, and classical formulae are employed. The bootstrap procedure, used to estimate empirically the distribution especially in presence of skewed distributions, is generally slightly biased, comparing to the theoretical quantile *Z* of the normal distribution. Such bias appears negligible especially for small standard deviations. Further investigations with resampling techniques may produce results for other location parameters.

*Department of Scienze Statistiche e*
*Matematiche “S. Vianelli”,*
*University of Palermo*

MASSIMO ATTANASIO
MARCO ENEA

*Ultrasuoni s.r.l., Catania, Italy*

LEONARDO RIZZO

## REFERENCES

U. ARENA, F. VIZZUTTI, G. CORTI, S. AMBU, C. STASI, S. BRESCI, S. MOSCARELLA, V. BODDI, A. PETRARCA, G. LAFFI, F. MARRA, M. PINZANI, (2008), *Acute viral hepatitis increases liver stiffness values measured by transient elastography*, "Hepatology", 47, pp. 380-384.

A.A. BRAVO, S.G. SHETH , S. CHOPRA, (2001), *Liver biopsy*, "New England Journal of Medicine", 344, pp. 495-500.

E. CAREY, W.D. CAREY, (2010), *Noninvasive tests for liver disease, fibrosis, and cirrhosis: is liver biopsy obsolete?*, "Cleveland Clinic Journal of Medicine", 77, 8, pp. 519-527.

L. CASTERA, (2009), *Acoustic radiation force impulse imaging: a new technology for the noninvasive assesment of liver fibrosis?*, "Journal of Gastrointestinal and Liver Diseases", 18, 4, pp. 411-2.

W.G. COCHRAN, (1977), *Sampling Techniques*, John Wiley & Sons, New York.

J. COHEN, (1968), *Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit*, "Psychological Bulletin", 70, 4, pp. 213-220.

M. FRIEDRICH-RUST, M.F. ONG, S. MARTENS, C. SARRAZIN, J. BOJUNGA, S. ZEUZEM, E. HERMANN, (2008), *Performance of transient elastography for the staging of liver fibrosis: a meta-analysis*, "Gastroenterology", 134, pp. 960-974.

J. HWANG, A. BISWAS, (2008), *Odds ratio for a single 2 x 2 table with correlated binomials for two margins*, "Statistical Methods and Applications", 17, pp. 483-497.

Q. MCNEMAR, (1947), *Note on the sampling of the difference between corrected proportions or percentages*, "Psychometrika", 12, pp. 153-157.

M.L. PALMERI, M.H. WANG, J.J. DAHL, K.D. FRINKLEY, K.R. NIGHTINGALE, (2008), *Quantifying hepatic shear modulus in vivo using acoustic radiation force*, "Ultrasound in Medicine and Biology", 34, pp. 546-558.

H.A. SALEH, A.H. ABU-RASHED (2007) *Liver biopsy remains the gold standard for evaluation of chronic hepatitis and fibrosis*, "Journal of Gastrointestinal and Liver Diseases", 16, 3, pp. 263-6.

L. SANDRIN, B. FOURQUET, J.M. HASQUENOPH, S. YON, C. FOURNIER, F. MAL, C. CHRISTIDIS, M. ZIOL, B. POULET, F. KAZEMI, M. BEAUGRAND, R. PALAU, (2003), *Transient elastography: a new noninvasive method for assessment of hepatic fibrosis*, "Ultrasound in Medicine and Biology", 29, 1705-1713.

G. SHIHA, ET AL., (2009), *Liver fibrosis: consensus recommendations of the asian pacific association for the study of the liver (apasl),* "Hepatology International", 3, 2, pp. 323-333.

M. SULLIVAN PEPE (2003*) The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press.

J.A. TALWALKAR, D.M. KURTZ, S.J. SCHOENLEBER, C.P. WEST , V.M. MONTORI, (2007), *Ultrasound-based transient elastography for the detection of hepatic fibrosis: systematic review and meta-analysis*, "Clinical Gastroenterology and Hepatology", 5, pp. 1214-20.

## SUMMARY

*Some issues concerning the statistical evaluation of a screening test: the arfi ultrasound case*

In this paper we analyze some issues concerning the statistical evaluation of a screening test for classification. The case study is ARFI, an ultrasound device recently introduced, and used for the evaluation of liver fibrosis. First, we present a simple statistical evaluation based on a novel index that compare two competitors with respect to a gold standard, and then we propose a procedure that determines a table with the "acceptable" number of measurements to get an "accurate" diagnosis using the ARFI device.