How Do Statistical Differences in Matrix-sensitive Magnetic Resonance Outcomes Translate Into Clinical Assignment Rules?

Richard G. Spencer, MD, PhD and **Nancy Pleshko, PhD**

Magnetic Resonance Imaging and Spectroscopy Section, National Institute on Aging, National Institutes of Health, Baltimore, MD (Dr. Spencer) and the Tissue Imaging and Spectroscopy Lab, Department of Bioengineering, Temple University, Philadelphia, PA (Dr. Pleshko)

There is an inherent difference in perspective between clinical research exploring matrix-sensitive MRI outcome measures and eventual application of these outcomes to individual patients. Basic science studies of group differences resulting from disease or treatment, or from differing demographics, for example, generally focus on the statistical significance of group mean differences. In contrast, the goal of a clinical measurement is to determine whether an individual patient belongs to a given group; this assignment is, in effect, a binary decision between normal and diseased.

Let us take an example from the literature in which enzymatic cartilage degradation is used to model the osteoarthritic process. These experiments are relatively straightforward, and the effects can be quite large; here we focus on T2 measurements as perhaps the most popular matrix-related magnetic resonance outcome. In one study, control cartilage exhibited a mean ± standard deviation (SD) $T2_{Ctrl}$ value of 55.0 ± 11.1 ms (n = 40), while after 18 hours of degradation with trypsin, values were significantly higher, with $T2_{Deg} = 66.5 \pm 10.8$ ms (n = 40).[1]

How do such results translate into the clinical viewpoint of detection of disease? Clearly, if the T2 values for two groups are very different, and if scatter within each group is small, then the assignment of a sample to control or degraded cartilage will be correspondingly more reliable. That is, the T2 measurement will exhibit a greater sensitivity (*SE*; accurate detection of degraded cartilage) and specificity (*SP*; accurate detection of intact control cartilage).

This can be formalized in several ways. First, one must specify how a sample's T2 measurement will be interpreted in terms of group assignment. One reasonable approach is to assign a new sample with a measured value $T2_{new}$ to the group whose mean value is closer to $T2_{new}$.[1] Thus, if the new sample has $T2_{new} = 61$ ms, it would be assigned to the degraded cartilage group because 61 ms is closer to 66.5 ms than it is to 55.0 ms. *SE* and *SP* can readily be defined within this framework. Consider two groups, a nondiseased control

group, *Ctrl*, and a group with disease, *Dis*, with means μ_{Ctrl} and μ_{Dis} for parameter p, and a measured value p_{new} from a patient, Pt. Then:

$$SE = \Pr\{|p_{new} - \mu_{Dis}| < |p_{new} - \mu_{Ctrl}| \mathrm{Pt} \in Dis\}$$

and

$$SP = \Pr\{|p_{new} - \mu_{Ctrl}| < |p_{new} - \mu_{Dis}| \mathrm{Pt} \in Ctrl\}$$

In words, the first of these definitions reads: *SE* is the probability (Pr) that, for a patient with the disease, p_{new} will be closer to the previously determined mean of the disease group than to the previously determined mean of the control group. That is, a diseased patient will be correctly classified as in fact having disease, based on the measurement of p_{new} in that patient. Similarly, the second expression may be read as: *SP* is the probability that, for a patient without the disease, p_{new} will be closer to the previously determined mean of the control group than to the previously determined mean of the disease group. That is, a nondiseased patient will be correctly classified based on p_{new}.

To mathematically translate group means and SDs to values for *SE* and *SP*, an underlying probability distribution must be assumed. A gaussian distribution of values within each group is a reasonable starting point. Given group means and SDs $\{\mu_{Ctrl}, \sigma_{Ctrl}\}$ and $\{\mu_{Dis}, \sigma_{Dis}\}$, and $\mu_{Dis} > \mu_{Ctrl}$, a straightforward calculation yields

$$SE = \Pr\{|p_{new} - \mu_{Dis}| < |p_{new} - \mu_{Ctrl}| \mathrm{Pt} \in Dis\}$$
$$= 1 - \Phi\left[\frac{(\mu_{Ctrl} - \mu_{Dis})}{(2 \bullet \sigma_{Dis})}\right]$$

and

$$SP = \Pr\{|p_{new} - \mu_{Ctrl}| < |p_{new} - \mu_{Dis}| \mathrm{Pt} \in Ctrl\}$$
$$= \Phi\left[\frac{(\mu_{Dis} - \mu_{Ctrl})}{(2 \bullet \sigma_{Ctrl})}\right]$$

where, in accordance with standard notation, $\Phi[z]$ is the cumulative probability distribution of the standard gaussian distribution, that is, a gaussian with mean $\mu_{Std} = 0$ and SD $\sigma_{Std} = 1$. The first of these expressions then states that *SE* equals the integral of the standard gaussian distribution between the integration limits $(\mu_{Ctl} - \mu_{Dis})/(2 \bullet \sigma_{Dis})$ and ∞; the second indicates that *SP* equals the integral of the standard gaussian distribution between the limits $-\infty$ and $(\mu_{Dis} - \mu_{Ctrl})/(2 \bullet \sigma_{Ctrl})$. Similar expressions are obtained for $\mu_{Ctrl} > \mu_{Dis}$.

With these expressions, the *SE* and *SP* of a clinical test, or decision rule, are obtained by simply plugging in the means and SDs for groups *Ctrl* and *Dis*. It is generally the case that the population means and SD are unknown, so one must instead use values from previous studies.

For the data presented above, that is, $T2_{Ctrl} = 55.0 \pm 11.1$ ms and $T2_{Deg} = 66.5 \pm 10.8$ ms, reflecting a statistically significant increase in T2 with degradation, the accuracy of the T2 measurement to detect whether a new sample belongs to the control group or to a group that underwent enzymatic degradation is rather poor, with $SE = 0.70$ and $SP = 0.70$. These values will further decrease if one accounts for random measurement error, which in effect broadens the underlying gaussian distributions and therefore yields even poorer test characteristics.

Let us now look at some of the highest-quality data available on matrix-sensitive magnetic resonance outcome measures for cartilage group differences in human subjects. It was reported that T2 measured at 3 Tesla was larger in subjects with osteoarthritis ($T2_{Dis} = 39.63 \pm 2.69$ ms; n = 10) compared to controls ($T2_{Ctrl} = 34.74 \pm 2.48$ ms; n = 10)($P = 0.001$).[2] Using the above equations, these values translate to $SE = 0.82$ and $SP = 0.84$ for determining whether a given patient has osteoarthritis based on a T2 measurement. Therefore, these highly statistically significant group differences translate into a clinical decision rule with only modest accuracy.

Further limitations are evident if one incorporates random measurement error; for example, if one (somewhat optimistically) assumes a random error of 2 ms in the T2 measurement, then $SE = 0.77$ and $SP = 0.78$. The corresponding values for a 4-ms random error are $SE = 0.70$ and $SP = 0.70$. These limitations stem from the narrow dynamic range of magnetic resonance parameters over clinical populations, resulting in parameter value overlap between groups. In contrast, the sensitivity and specificity of magnetic resonance diffusion-weighted imaging for acute ischemic stroke, when performed >12 hours after the event, have been reported to be 92% and 97%, respectively.[3]

The fact that statistically significant group outcome measures do not necessarily translate into useful clinical outcome measures currently limits the utility of quantitative cartilage matrix–sensitive MRI in clinical decision making. The optimal tactic to create more clinically meaningful tests from basic science studies is not obvious. One promising approach is to apply multivariate statistical techniques, in which changes in several magnetic resonance parameters can be combined to improve SE and SP over values exhibited by any one parameter individually.[4,5] This has the advantage of not requiring new acquisition protocols, or physics and hardware developments, although it does necessitate a change in perspective. The implementation of these somewhat complex analytic approaches will require considerable exploration but may hold a great deal of promise.

Acknowledgments

References

1. Lin PC, Reiter DA, Spencer RG. Sensitivity and specificity of univariate MRI analysis of experimentally degraded cartilage. Magn Reson Med. 2009; 62(5):1311–1318. [PubMed: 19705467]

2. Li X, Benjamin Ma C, Link TM, et al. In vivo T(1rho) and T(2) mapping of articular cartilage in osteoarthritis of the knee using 3 T MRI. Osteoarthritis Cartilage. 2007; 15(7):789–797. [PubMed: 17307365]

3. Chalela JA, Kidwell CS, Nentwich LM, et al. Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: A prospective comparison. Lancet. 2007; 369(9558):293–298. [PubMed: 17258669]

4. Lin PC, Irrechukwu O, Roque R, Hancock B, Fishbein KW, Spencer RG. Multivariate analysis of cartilage degradation using the support vector machine algorithm. Magn Reson Med. 2012; 67(6): 1815–1826. [PubMed: 22179972]

5. Lin PC, Reiter DA, Spencer RG. Classification of degraded cartilage through multiparametric MRI analysis. J Magn Reson. 2009; 201(1):61–71. [PubMed: 19762258]