Controversies in Multiple Sclerosis

Modeling the course and outcomes of multiple sclerosis is statistical twaddle—Yes

Bruce V Taylor

Our ability to modify the outcomes of multiple sclerosis (MS) is reliant on our ability to measure disability and change in disability reliably, reproducibly, precisely, and ideally at all stages of MS. We cannot determine whether interventions, pharmacological or otherwise, alter the natural history of MS if we cannot model the natural history accurately at the individual and group level. These apparently simple metrics, that is, whether the person with MS getting worse, staying the same or improving, have proven elusive, and to date, this problem has not been solved. What MS desperately requires is such a metric of severity and change. Since at least the 1930s, various MS diagnostic criteria, MS clinical course definitions and clinical and para-clinical disability measures have been proposed. The fact that none of these measures has survived the test of time without modification, or significant qualification, attests to the complexity of the disorder.

Disability defined by a committee

One of the most striking things evident in MS is that we do not have clear diagnostic or biological markers and rely on consensus criteria to define the disease and its clinical course. Consequently, we are left trying to assess disability and clinical course in a disease whose metrics are defined by a committee. Perhaps the most difficult measure of all is defining and measuring progression of disability. Since Kurtzke first proposed the Disability Status Scale (DSS) and the subsequent Expanded DSS (EDSS), it has been referred to as the gold standard of disability measure. The EDSS is widely used and the European Medicines Agency (EMA) recommends it as a scale for use in clinical trials of MS. It has many useful attributes, in that it is well-recognised, widely utilised and relatively easy to measure. Where it has well-documented shortcomings is that it is non-linear, has a population bimodal distribution with peaks around EDSS 3 and 6 and the time to progression between intervals is not uniform. It also measures different aspects of disability at different points along the scale. There are also issues with inter- and intra-rater reliability and what constitutes a significant change in disability. It also has poor to, at best, fair correlation with magnetic resonance imaging (MRI) measures and other para-clinical measures of MS progression such as ocular coherence tomography, leading to the recent definition of clinical-radiological-ophthalmic paradox (CROP) in MS where these measures of progression do not correlate at all well.

More recently, the MS functional composite (MSFC) score has been developed supposedly to overcome the difficulties of the EDSS. The MSFC has become widely used in clinical trials but not widely in clinical practice. It, like the EDSS, has thrown up problems with reproducibility, lack of correlation with other measures, difficulty in defining a reference population, poor patient acceptance and also a noticeable learning effect. Similar difficulties are encountered for clinical scales such as the Scripps Score and the Guys Neurological disability score.

Poor measures mean poor prediction

So if we cannot measure an individual’s level of disability reliably and reproducibly and cannot reliably measure change, how can we define a reliable measure of efficacy in clinical trials that reflects the known vagaries of the disease? This question has resulted in much conjecture and assessment particularly in the field of clinical trials of MS therapy. The ideal study would be long-term compared to placebo, of large size and would use a number of independent validated end-points to judge efficacy. With the switch in emphasis to prevention of and treatment of progressive MS, the importance of outcome measures that provide sensitive and specific markers of disease progression has become even more important. Much has been made of the use of MRI, particularly change in MRI brain volume metrics, as the potential next gold standard of MS progression. However, statistically, there are a number of issues, not the least being the difficulty of reproducibility, with the small relative changes seen over modest time intervals, that all lead to significant coefficients of variation (CVs) for every measure meaning that, at the individual level, it is difficult to use MRI...
volumetrics to measure progression. At the group level, there is probably a greater chance of using brain atrophy as a surrogate marker of MS progression. However, again, the reliability of this measure over the duration of a 2-year clinical trial has been questioned.

Over the last few years, the concept of sustained disability progression over 3 or 6 months has been suggested as an important outcome in clinical trials; however, recent work has suggested that this time interval should be at least 12 months. This work also suggested that the minimum level of difference in serial EDSS measurements that should be considered significant is 1 EDSS point. As the average intra-rater reproducibility of the EDSS, particularly within the lower half of the EDSS scale is >1 point, lower levels of EDSS change are not likely to be reproducible and therefore not statistically valid. Some authors have advocated that 1.5 EDSS points is the minimum that can be reliably detected and, particularly for EDSS scores below 5.5, a 1.5 point change would appear to be significantly more robust as an outcome metric above EDSS 5.5, the figure may be lower. Additionally, sustained EDSS progression may occur because of non-MS specific disease processes such as de-conditioning, weight gain and co-morbidities. Another major concern is that disability does not progress at the same rate, even within clinical studies, with disability progressing faster in the first year than the second year, often regressing to the mean.

An additional interest over the last few years has been modeling clinical course from baseline measures; that is predicting outcomes from early disease stages, using mathematical models that take into account multiple parameters noted at baseline. These scores have reasonable sensitivity in predicting outcomes = 80%. However, these scores BREMS and BREMSO again reflect the potential issues with statistical modeling; significance is seen for differences between those in the first and fourth quartiles of severity as measured by the multiple sclerosis severity score (MSSS), but in the BREMSO score 15% of those with severe MS had a low BREMSO score and 16% with mild MS had a high BREMSO score. For those in the middle two quartiles of severity, the scores do not discriminate between good and poor prognosis with any reliability. This suggests that statistical modeling of clinical course may be possible for those with extremes of phenotype (where generally we already have a good idea from clinical experience what will happen) but not useful for the majority of people with MS.

**Needed a good marker of progression**

So many have sought to find statistical significance in changes in MS clinical and para-clinical metrics and predict the course of MS using baseline parameters, within reality at best modest success. One of the principle issues, in my opinion, is not understanding the significant natural variation that can occur in measurement of disability, and not taking into account the vagaries of the measures being used. If you are trying to assess progression, you need a gold standard or end-point and, short of death, there is no good marker. The changes may come on gradually in MS, may reverse and may be driven by extraneous factors. We as MS clinicians need to be aware of these when assessing any claim made of efficacy particularly when the changes are small. The chance of a small but statistically significant change in a marker of MS progression being due to measurement error is high and cannot, at the present stage, be avoided.

Composite measures are also very tricky to assess particularly as they include a number of measures all with their own CVs. The significance of these measurement variances should be taken into account when judging what a significant change is and what magnitude of p value should be accepted from these measures. Certainly, with a composite measure, multiple testing should be considered and at a minimum the number of tests included should be adjusted for.

So when we as clinicians are presented with the results of a clinical study in MS or are involved in the design of studies, we need to carefully look at the outcome with the knowledge of what it is built on. If the study is not large, does not have a significant follow-up period, is not adjusted for multiple testing and does not specify how time varying variables are adjusted for, the results are likely to be ‘statistical twaddle’ and should be interpreted as such. We cannot accept small, just significant, composite end-points; they are meaningless and should be confined to the dustbin of failed studies. Similar arguments can be made when considering predictive scores used to model MS clinical course; they essentially tell us what we already know but do not provide information that is useful for those cases in the middle who still have a huge variability in their outcomes.

We have a greater challenge than dealing with statistical twaddle; we need better end-points, we need a gold standard measure of disability and we need it now. What we should not do is give up; we need to continue measuring disability to the best of our abilities. The EDSS is the best measure we currently have, but we need to understand its metrics better and how it works in large real-world settings, and with refinement, it may continue to provide us with outputs that we understand and can apply clinically.
Modeling the course and outcomes of MS is statistical twaddle—No

Roberto Bergamaschi and Cristina Montomoli

Multiple sclerosis (MS) is classically described as an unpredictable disease in which the prognosis varies greatly, with some patients remaining asymptomatic for a long time and other patients becoming severely disabled in a short time. Thus, the ambition of modeling the course and outcomes of MS could be judged “statistical twaddle.”

Nevertheless, a major challenge in MS management is to predict the long-term evolution of individual...