Multiple Sclerosis and Related Disorders 18 (2017) 109–116

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

journal homepage: www.elsevier.com/locate/msard

Review article

# Comparing the efficacy of disease-modifying therapies in multiple sclerosis

Dimos D. Mitsikostas[a,1], Douglas S. Goodin[b,*,1]

[a] National and Kapodistrian University of Athens, 1st Department of Neurology, Aeginition Hospital, Athens, Greece
[b] University of California, San Francisco, Department of Neurology, 505 Parnassus Ave, Suite M-794, San Francisco, CA 94143-0114, USA

ARTICLE INFO

*Keywords:*
Multiple sclerosis
Therapy
Efficacy
Safety
Relative risk
Absolute risk reduction
Number needed to treat
Cross-trial
Evidence-based
Comparison

ABSTRACT

Establishing the relative efficacy and safety of the different disease modifying therapies (DMTs) in multiple sclerosis (MS) is critical to the choice of agent that clinicians recommend for individual MS patients. The best evidence for the relative efficacy of the different DMTs comes from head-to-head randomized clinical trials (RCTs). Understanding that outcome-measures with the best established validity are the relapse rate and the actual (not the "confirmed") change in the extended disability status scale (EDSS), we conclude from these head-to-head RCTs that interferon-beta (IFNβ) given subcutaneously multiple times per week (either IFNβ-1b or IFNβ-1a) and glatiramer acetate (GA) are about equivalent in terms of efficacy and that both of these agents, as well as many of the other DMTs, are superior to weekly intramuscular IFNβ-1a. Nevertheless, as ever-newer agents with novel mechanisms of action are brought to the marketplace, such direct head-to-head trials are becoming increasingly impractical, raising the need for alternative methods to draw reasonable inferences from less rigorous clinical data. One possible approach to judging comparative efficacy is to make comparisons across clinical trials using the complimentary analytic methods of calculating both the relative risk/rate and the absolute risk/rate reductions. A consideration and application of this analytic approach is undertaken here. It is only with an understanding of the safety and efficacy of the different agents that we can select, together with the patient, the right agent for the right person.

## 1. Introduction

The first modern disease-modifying therapy for MS – IFNβ-1b – was introduced in 1993 (The IFNB Multiple Sclerosis Study Group, 1993; Paty et al., 1993). Shortly thereafter, two different formulations of IFNβ-1a and a novel treatment – GA – also became available (Johnson et al., 1995; Jacobs et al., 1996; Simon et al., 1996; PRISMS Prevention of Relapses and Disability by Interferon beta-1a Subcutaneously in Multiple Sclerosis Study Group, 1998; Li et al., 1999). For more than a decade, these DMTs were the only proven-effective MS-treatments available. Moreover, these four therapies represented only two approaches to the treatment of MS (i.e., IFNβ or GA). However, beginning in 2006 with the introduction of natalizumab (NTZ), both the number of available DMTs, and the number of therapeutic approaches has increased dramatically (Polman et al., 2006; Agoropoulou et al., 2010; Calabresi et al., 2014; O'Connor et al., 2011; Confavreux et al., 2014; Gold et al., 2012; Fox et al., 2012; Hauser et al., 2008; Cohen et al., 2012; Coles et al., 2012; Gold et al., 2013; Hauser et al.). We now have 14 agents in nine different classes approved for use in MS (The IFNB Multiple Sclerosis Study Group, 1993; Paty et al., 1993; Johnson et al., 1995; Jacobs et al., 1996; Simon et al., 1996; PRISMS Prevention of Relapses and Disability by Interferon beta-1a Subcutaneously in Multiple Sclerosis Study Group, 1998; Li et al., 1999; Polman et al., 2006; Agoropoulou et al., 2010; Calabresi et al., 2014; O'Connor et al., 2011; Confavreux et al., 2014; Gold et al., 2012; Fox et al., 2012; Hauser et al., 2008; Cohen et al., 2012; Coles et al., 2012; Gold et al., 2013; Hauser et al.) and several others either awaiting approval or in earlier developmental-stages. This glut of available-options has created, for practitioners, a dilemma about which agent to use, in which patient, and under what circumstances. To make such a decision, however, requires an accurate comparison of the different therapeutic agents with regard to their efficacy, safety, tolerability, and convenience. In a perfect world, the data to make these comparisons would come from well-designed, head-to-head RCTs.

Nevertheless, in the real world, obtaining this type of rigorous clinical data is neither possible nor is it, necessarily, desirable. For example, comparing the nine different classes of therapeutic agents requires either one extremely-large, likely infeasible, RCT, or, if these therapies were to be compared two at-a-time, the conduct of 36 separate RCTs to assess the relative efficacy. Moreover, even then, the

* Corresponding author.
*E-mail address:* douglas.goodin@ucsf.edu (D.S. Goodin).
[1] The authors contributed equally to this work.

http://dx.doi.org/10.1016/j.msard.2017.08.003
Received 25 September 2016; Received in revised form 6 June 2017; Accepted 6 August 2017
2211-0348/ Published by Elsevier B.V.

results would likely prove difficult to interpret. To undertake such a project, either by a government agency or the pharmaceutical industry, would be prohibitively expensive, extremely time-consuming, and would almost certainly delay the development of other novel therapeutic approaches. Consequently, practitioners both now and in the future will need to base their therapeutic decisions upon less rigorous clinical evidence.

Broadly, three general approaches can be used to compare the different DMT agents with regard to their efficacy, safety, and tolerability. The first is to use the available head-to-head data and to draw, from these RCTs, inferences about agents that have not been directly-compared. The second is to use the pivotal-RCT data and to make comparisons across the different clinical trials. And the third is to use the post-RCT clinical-experience with the different agents. Each of these approaches has advantages and disadvantages. We consider here the merits and limitations of these approaches for making such comparisons in the absence of complete head-to-head data.

## 2. Evaluating outcome

Regardless of the comparison-method, all outcomes measured during the RCT are, at best, only surrogates for the long-term outcome (i.e., fixed unremitting disability – physical or cognitive) that we hope to prevent or postpone with our therapies. Therefore, we need to establish which of these short-term RCT-outcomes actually correlates with long-term disability. To validate a surrogate requires more than simply establishing a correlation (Prentice, 1989). Nevertheless, without a correlation, the outcome-measure cannot be valid. For example, just labeling something a "disability" measure doesn't mean that it actually measures disability. In RCTs, short-term disability is generally assessed using the 10-point EDSS scale, developed by Kurtzke in the 1950s and subsequently-modified (Kurtzke, 1955, 1983). Nevertheless, despite this long tradition, even EDSS changes require a correlation with long-term outcome to be considered potentially valid short-term outcome-measures for use in RCTs.

To establish these correlations, requires that long-term data be acquired following the conclusion of the RCT, that patient-ascertainment be as complete as possible, and that the measure used to assess fixed long-term disability be as unambiguous as possible. For physical disability, examples of such unambiguous or "hard" outcome-measures include unremitting-EDSS $\geq 6$, unremitting-EDSS $\geq 7$, and death due to MS. Unlike other "softer" disability outcomes such as a change on the EDSS which is "confirmed" 3 months later, these "hard" outcomes cannot revert back to baseline over time and are final. Consequently, these "hard" outcomes are appropriately analyzed using survival or time-to-outcome methods. Conversion to secondary-progressive (SP) MS can also be a "hard" outcome of advancing disease-severity but pinpointing the time of this transition can be difficult, at least contemporaneously. Finally, although cognitive disability is of critical importance to both patients and families, "hard" long-term cognitive disability outcomes are not currently available.

There has only been limited long-term follow-up (LTF) experience from which to draw inferences. For example, in some LTF studies, either case ascertainment is so low (~40%) or the entry criteria are such that the reported findings, almost certainly, are contaminated by substantial selection-bias (Goodin, 2004, 2013; Bermel et al., 2010; Ford et al., 2010; Shirani et al., 2012; Goodin et al., 2012a). In addition, several LTF studies have evaluated only "soft" disability outcomes or, when "hard" outcomes were assessed, the relationship (i.e., correlation) between these outcomes and the short-term RCT-measures has not been explored (Bermel et al., 2010; Ford et al., 2010; Goodin, 2013).

Two LTF studies stand out (Liu and Blumhardt, 2000; Goodin et al., 2012b). The first was by Liu and Blumhardt (Liu and Blumhardt, 2000), who found that, regardless of how "confirmed" disability progression was defined, approximately half of the progressed-patients reverted to non-progressed status by the end of the RCT (Table 1). Thus, these particular "confirmed-disability" outcome-measures could not even be validated over the course of the RCT. Similarly, Goodin and colleagues (Kappos et al., 2006) found that the 1-point EDSS progression, confirmed at three months, was only weakly correlated with 16-year long-term disability and accounted for only about a sixth of the variance as that which was accounted for by the actual change in EDSS over the entire course of the RCT (Table 2). Moreover, using a regression analysis with stepwise elimination, neither the 1-point "confirmed" EDSS change nor any of the "on RCT" MRI outcomes were retained in the final model (Table 3).

**Table 1**
Confirmed vs. sustained EDSS change.[a]

| Definition | | EDSS Change | | |
|---|---|---|---|---|
| | | Confirmed | Sustained | PPV |
| 1 Point | 3 Months | 32·2% | 15·3% | 0.48 |
| 1 Point | 6 Months | 21·4% | 14·1% | 0.67 |
| 2 Points | 3 Months | 12·1% | 6·4% | 0.53 |
| 2 Points | 6 Months | 9·3% | 5·1% | 0.55 |

Sustained Change = remains changed at end of RCT.
PPV = positive predictive value.
Data from Liu and Blumhardt (Goodin, 2013).
[a] Confirmed Change = meets the definition of progression.

These findings suggest that when comparing the efficacy of different DMTs, attention should focus on the clinical measures of attack rate and the EDSS change over the entire RCT, and not on either the effect of a drug on the "confirmed" EDSS progression or the magnitude of an agent's MRI impact. Such a conclusion also has important implications for the new, and currently popular, outcome of "no evidence of disease activity" or NEDA (Nixon et al., 2014; Rotstein et al., 2015). Like any other short-term measure, NEDA needs to be validated by its correlation with long-term disability. Nevertheless, because of its sensitivity to MRI-outcomes, it seems likely that, similar to other "on RCT" MRI measures, NEDA may actually contribute very little to the prediction of long-term disability (Table 3).

## 3. Head-to-head evidence

Only a limited amount of head-to-head RCT data is available. For example, head-to-head studies of subcutaneous IFNβ-1b or subcutaneous IFNβ-1a given multiple times per week (Durelli et al., 2002; Panitch et al., 2002), GA (Lublin et al., 2013), fingolimod (FGL) (Cohen et al., 2010), NTZ (Rudick et al., 2006), and daclizumab (DCL) (Kappos et al., 2014) have each demonstrated superiority on relapse rate and several MRI measures when compared to weekly intramuscular IFNβ-1a. By contrast, head-to-head evidence comparing GA either to subcutaneous IFNβ-1a or to IFNβ-1b showed no difference between the treatments with regard to either relapse rate or disability (Mikol et al., 2008; O'Connor et al., 2009). Although, on some MRI measures, either formulation of IFNβ was superior to GA, these measures were inconsistent between trials (Mikol et al., 2008; O'Connor et al., 2009).

Both the FGL trial (Cohen et al., 2010) and the NTZ trial (Rudick et al., 2006) included patients who might be construed as "treatment failures" on weekly IFNβ-1a and who, nevertheless, were subsequently randomized to receive the failed-therapy. Consequently, the superiority of FGL or NTZ therapy over weekly IFNβ-1a may have been exaggerated. Despite this, however, the superiority of FGL over weekly IFNβ-1a was actually greater in treatment-naïve patients than it was in patients who were possible "treatment failures" (Cohen et al., 2010). In the NTZ trial (Rudick et al., 2006), the superiority of NTZ over weekly IFNβ-1a was of such a magnitude that this theoretical concern is likely mitigated (Rudick et al., 2006). Thus, even these trials seem to establish that FGL and NTZ are superior to weekly intramuscular IFNβ-1a. Despite these conclusions, however, no insight is provided by these