Overcoming obstacles in clinical trial design: The MS experience

Irene Cortese
Neuroimmunology Branch Clinical Group, National Institute of Neurological Diseases and Stroke, NIH

1. Introduction

Konnichiwa. I want to thank the organizers, and especially Professor Matsumoto, for giving me the opportunity to be here today to share with you some of the challenges that we have faced in the development of clinical trials in multiple sclerosis (MS), and how we have attempted to solve these challenges.

2. General rule and goal of clinical trials

I'm going to start with just a general statement that I think we can all agree on which is that appropriate study design needs to include sufficient sample size and statistical power and control of bias to allow for a meaningful interpretation of results.

I will go further and say that the ultimate goal of pivotal clinical trials, that is, trials conducted to obtain regulatory approval for use, is to provide definitive evidence regarding the benefit-to-risk profile of the experimental intervention respective to placebo or an existing standard of care intervention. Since it is the risk and benefit that must both be assessed, a clinically meaningful endpoint should be used.

3. General problem in clinical trials

The problem, however, is when this clinically meaningful endpoint is uncommon. In this case, we have to increase our sample size or increase the trial duration greatly in order to maintain statistical power. This has a significant implication for cost, as well as for feasibility of the trial itself. There's an additional problem in the particular circumstance of rare diseases where raising the sample size may not be an option at all.

There are different ways we try to overcome this problem. We can select endpoints that are more easily measured. And we can use trial designs that try to maximize efficiency and maximize the use of resources.

4. Surrogate endpoint and biomarker: Definition

One such approach is that of using a surrogate endpoint. A surrogate endpoint has been defined as a laboratory measurement or physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives. It may be related to the clinical

true endpoint, but the relationship between the two may not be direct.

Surrogate endpoint is different from a biomarker. A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic function or pathogenic processes or pharmacological responses to a therapeutic intervention.

5. Biomarkers vs. surrogate endpoints

The relationship between a biomarker and surrogate can be stated to be that surrogate endpoints are a subset of biomarkers. A biological biomarker is a candidate for a surrogate endpoint if it is expected to predict clinical benefit – either of harm, lack of benefit or lack of harm – based on epidemiological, therapeutic, pathophysiologic or other scientific evidence.

6. Surrogate endpoints: Characteristics and advantages

Surrogate endpoints can replace a distal endpoint with a more proximal one. They can generally be measured earlier. They can be measured more easily or frequently. And they can be measured with higher precision. They may be less affected by other treatment modalities. They generally can reduce sample size requirements and lead to the possibility of faster decision making.

In general we can say that advantages of the surrogate endpoint are that the study becomes simpler, shorter and less expensive.

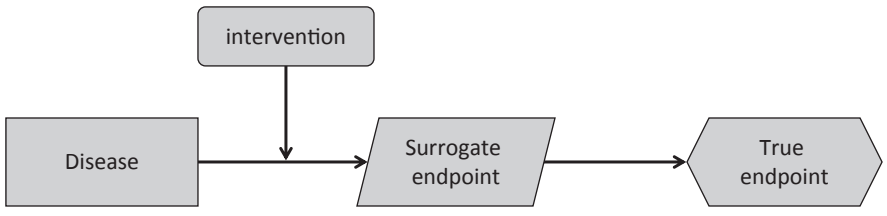
For these reasons, surrogate endpoints are desirable in early clinical trials and also in trials of rare diseases.

7. Ideal surrogate endpoint (Fig. 1)

We have to remember though that simply the fact that a biomarker is correlated with an endpoint does not make it a good surrogate endpoint.

We can describe an ideal surrogate endpoint as being an endpoint that sits along the causal pathway of the true endpoint, and therefore can reflect this pathogenic process. Even in this ideal setting however, we have to be cautious because a surrogate endpoint can be misleading in terms of the magnitude of a treatment effect. For example, we might overestimate a treatment intervention if the surrogate endpoint is statistically significant however the treatment effect is not of sufficient size or magnitude or sufficient duration to actually impact the true endpoint of interest. On the other hand, we might underestimate a true treatment effect if there is "noise in the system" which does not allow us to adequately measure this surrogate endpoint.

Fig. 1 Ideal surrogate endpoint



Gupta K, Gupta J, Singh S. Surrogate endpoints: How reliable are they? *J Clin Res Best Pract.* 2010 ; 6(5).

8. Cardiac Arrhythmia Suppression Trial (CAST)

Probably the most cited example of a time when a surrogate endpoint led to false information, or was misleading and actually led to clinical harm in patients, was the Cardiac Arrhythmia Suppression Trial or CAST. It was known that ventricular arrhythmia is associated with a four-fold increase in the risk of death following myocardial infarction. So two drugs – Encainide and Flecainide – were found to suppress ventricular arrhythmia effectively by ECG, and these two drugs were approved for use by the FDA based on this surrogate endpoint. The drugs were used widely in the post-myocardial period and ventricular arrhythmias were indeed suppressed. However, eventually a confirmatory study – the CAST study – was performed using the true clinical outcome of overall survival after myocardial infarction. It was very difficult to recruit for this study because the drugs were now already FDA approved, and there was much reluctance on the part of physicians who felt that it was unethical to withhold this treatment which was surely effective in reducing mortality. But eventually this study was done, and this showed surprisingly that these two drugs actually led to a three-fold increase in death over placebo. So how can this happen?

9. How surrogate endpoints can mislead

There are a few ways that we can imagine that a surrogate endpoint might mislead us. One possibility is that the surrogate endpoint is not actually on the causal pathway to the true endpoint, and therefore, our intervention can affect the surrogate endpoint but does not impact the true endpoint at all.

Another possibility is that the surrogate endpoint is indeed on a causal pathway to the true endpoint, but that there are multiple causal pathways. The disease is multi-factorial. So in this event, we may overestimate our treatment effect if we are only measuring the effect on this causal pathway.

Another possibility is that our surrogate endpoint is on a causal pathway, but not the one that is actually impacted by our intervention. In this case, if we use this surrogate endpoint to estimate the effect of our intervention, we may underestimate a possible true effect and we might not appreciate that this intervention could actually be helpful.

Another possibility is that the intervention itself can have a direct harmful effect on our true endpoint. Or alternatively the intervention could actually have a positive effect on an additional causal pathway that is not recognized. In both these cases, we may have an overestimation of the overall benefit of our drug or an underestimation of the treatment effect of our drug. So obviously, it becomes very clear that the validation of surrogate endpoints is very important.

10. Prentice Criteria: Validation of surrogate endpoints

Probably the most cited proposed criteria for the validation of surrogate endpoints are the Prentice Criteria. I have these written in the mathematical formula but I will use the paraphrasing which is more simple. The first postulate is that there is a significant treatment effect on the true endpoint. The second is that there must be a significant treatment effect on the surrogate endpoint. The third is that there must be a significant correlation between the surrogate and the true endpoint. And the last is that the effect of the treatment on the clinical endpoint must disappear or become

non-significant when adjusting for the surrogate – meaning that the whole of the treatment effect can be explained by the treatment effect on the surrogate.

11. Caveats with surrogate endpoints

Just a word of caution – because there are no hard rules as to how a surrogate endpoint is validated, as scientific readers and also as patient care providers, we must scrutinize studies very carefully when they are relying on a surrogate as the primary endpoint.

Also, we have to remember a few caveats. First of all, that validating a surrogate for one population or one intervention or one clinical outcome does not imply that the surrogate is valid for another patient population or other intervention or situation. In addition, we have to remember that a valid surrogate for benefits may not be a valid surrogate for the harms associated with a particular intervention.

12. FDA Fast Track Development Program

So in the end, what it comes down to is, proper validation of a surrogate requires in-depth understanding of the causal disease pathways as well as the intervention's intended and unintended mechanisms of action. This is the only way that we can truly validate a surrogate endpoint.

In 1992, on the heels of the AIDS epidemic in the United States, the FDA developed what is known as the Fast Track Development Program. This program was designed in order to facilitate the development and expedite the review and approval of new drugs that are specifically for serious or life-threatening conditions or that may meet unmet medical conditions – conditions for which there is no treatment.

13. 21 CFR 314, Sub-Part H – Accelerated Approval

This accelerated approval is contained in the Code of Federal Regulations (CFR) Sub-part H. What this rule basically allows is FDA approval or regulatory approval of medications under this process based on surrogate or non-ultimate clinical endpoints. However, the regulation also requires that post-marketing data be acquired in order to validate that surrogate endpoint and to demonstrate that there is indeed a clinical benefit that was anticipated based on that surrogate endpoint. This rule and this process have been somewhat controversial. Although showing biological effect, approving a medication based on a surrogate endpoint does not evaluate the full risk and benefit as we saw with the CAST trial. And once a drug is on the market, it can be very difficult to acquire this data that is necessary in order to formally and finally validate that trial.

14. Early clinical trials and small clinical trials

But when talking about early clinical trials, what we can say is that surrogate endpoints are very useful for demonstrating biological effect (e.g. in phase 2 trial) and for determining whether an intervention is promising enough to proceed to a large and definitive trial.

I just want to add that small clinical trials have some inherent difficulties. They are more prone to variability and may only be adequately powered to detect large intervention effects as a consequence. Therefore, the importance of adequate study planning is magnified in small clinical trials.

You see here a scheme that shows a number of different types of trial design. In particular, we have some trial designs that are especially suited for small clinical trials. The characteristic of these trials is that they try to maximize the information that can be gained from a single patient.

15. Neuroimmunology Branch, NIH

The Neuroimmunology Branch at the NIH was established in 1975, and as the name suggests, has a focus on neuroimmunological diseases. Since its inception, this branch of the NIH has been conducting disease-specific natural history studies in order to understand the biology of these diseases, and to develop databases to give access to researchers in and outside of the NIH to this valuable information; to identify, develop and validate novel biomarkers as well as surrogate endpoints, and also to explore the use of novel clinical trial design and statistical methods.

16. Multiple sclerosis

I said that we focus on neuroimmunological diseases, but the bulk of the research we do is in multiple sclerosis. Multiple sclerosis is a complex disease. It is multi-factorial. It is characterized by a balance between 3 different pathophysiological mechanisms – inflammation, demyelination, and neurodegeneration. It affects about one in a hundred Americans but the sub-type of primary progressive MS affects about one in 8,000 Americans which probably approaches more the incidence in Japan. There is treatment for relapsing, remitting MS which is the dominant form of MS, the most common form of MS. But this treatment is expensive. It's only partially effective and it often requires constant changes and modifications. So the bottomline is that for this disease there is no cure, there is no prevention, and there isn't even a known cause.

In 1993, the first drug for MS – Betaseron – was approved for the treatment of relapsing, remitting multiple sclerosis. This approval was based on the demonstration that Betaseron was able to suppress the frequency of relapses which is the major clinical manifestation of this disease.

Around this time, MRI was also gaining importance in the study of multiple sclerosis. It was already clearly important for the diagnosis of MS and it was beginning to gain importance also as a surrogate marker for tracking response to treatment. And it was recognized that the lesions that are typical of multiple sclerosis increase over time so that new lesions develop and the old ones persist.

17. MRI as a surrogate market of drug efficacy in MS

What was also noted was that when a new lesion develops such as this one, in its initial phase, it is contrast enhancing, meaning that it takes up gadolinium by MRI. This contrast enhancement is transient and it denotes active inflammation.

From natural history studies that were performed outside of NIH but also largely at NIH, what was seen when scanning patients every month by MRI is that the frequency of these contrast enhancing lesions (CEL) was far greater than the clinical relapses. In fact, the ratio was about five to one or even ten to one in some patients. This clearly suggested that contrast enhancing lesions in particular might be a very useful surrogate endpoint in clinical trials in MS because we are replacing a surrogate endpoint (clinical relapses) which is relatively infrequent (maybe one or two per year at the most) with a very common and frequent event that

could then shorten the lead time for screening potential candidate therapies considerably.

18. IFNβ-1b MS Study Group: Subgroup analysis (n=327) (Fig. 2)

So MRI was proposed as a surrogate marker for drug efficacy in MS, and it was in fact used in a planned sub-group analysis in that first pivotal trial for the approval of betaserone. This was done in a sub-set of patients who had frequent imaging performed in order to be able to make a determination of the effect of betaserone on contrast enhancing lesions. Going back to the model from before, what we can see is that using subclinical inflammation and the surrogate endpoint of contrast enhancing lesions, interferon beta was able to suppress contrast enhancing lesions and this was directly related to the suppression of relapses in this disease.

So what happened at NIH was there was the realization that this surrogate endpoint, which already could permit to screen therapies more rapidly, the efficiency could be improved even further if one used a different trial design.

19. MRI as a surrogate market in MS: Trial design considerations

Using a parallel group design (which is the design that was used in those original studies for the approval of betaserone) in order to achieve 80 percent power to detect the 60 percent decrease in contrast enhancing lesions, we would need two cohorts of 40 patients each and a period of six months. Using instead a single crossover design in which there is a 6-month run-in pre-treatment, followed by a 6-month period of treatment, we could achieve the same goal by using only 10 to 12 patients.

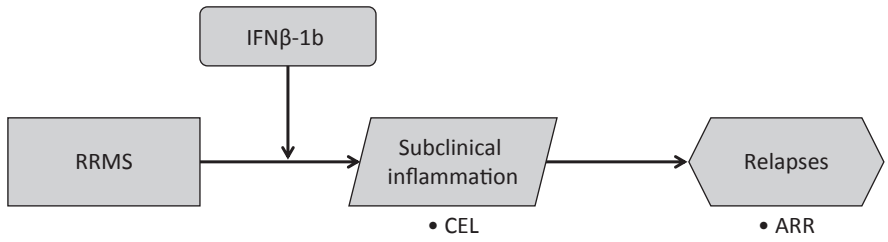
This design and this outcome was then successfully used and it was shown that interferon beta could in fact suppress contrast enhancing lesion formation using a group of just 14 patients. This was demonstrated clearly in this case. What was subsequently noted was that it was possible to use a shorter run-in period of just two months thereby significantly shortening the lead time further.

20. Model effectively used at NIH

This model was effectively used at NIH to screen for various candidate therapies for relapsing-remitting MS. The examples that you see here are for the clinical trials for APL (Altered Peptide Ligand) and for Roliprom which is a PD4 inhibitor in MS. Both of these trials were actually negative, meaning that they actually increased the formation of contrast enhancing lesions, as seen here graphically and seen here by MRI.

But this trial design and using the surrogate endpoint allowed to demonstrate that these thera-

Fig. 2 IFNβ-1b MS Study Group: Subgroup analysis (n=327)



peutic interventions would not be successful and ought to not move forward and so this was demonstrated in a much smaller number of people and in a much shorter time frame than could have been previously demonstrated.

21. Daclizumab for treatment of RRMS (relapsing remitting multiple sclerosis)

Not all the trials were negative. This same design was used also to demonstrate instead the efficacy of daclizumab which is an anti-CD25 monoclonal antibody for the treatment of relapsing remitting MS. As you can see here, the contrast enhancing lesions went down dramatically in these patients as did the formation of new lesions and their volume. This was also associated with improvement of clinical endpoints measuring disability.

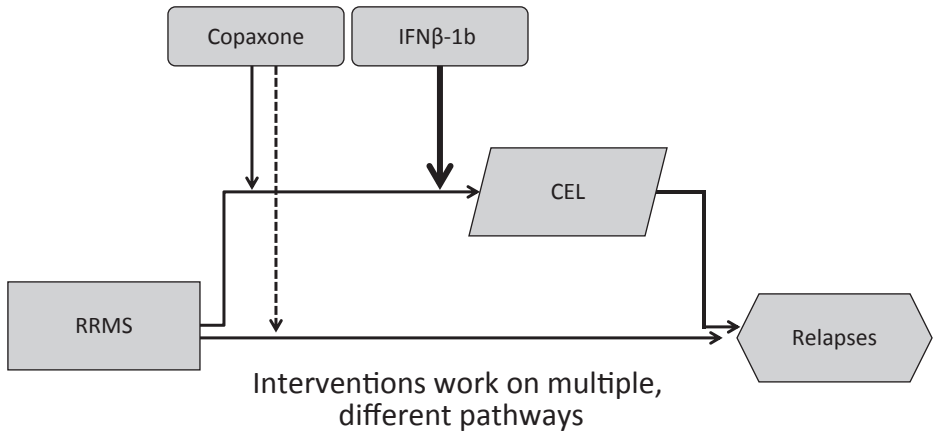
22. Surrogate markers are pathway specific

But I want to make another point here which is this – that surrogate markers are pathway specific. By this what I mean is that a surrogate for one population or one disease or one intervention cannot necessarily be applied to another group or another intervention. And this is shown in the results of the CombiRx study, which was a multicenter study of which the biomarker sub-study was actually carried out at NIH. In this study, there were 3 cohorts of patients. One cohort treated with interferon. One cohort treated with glatiramer acetate or Copaxone®. And another cohort treated with a combination of those two therapies. When we looked at clinical measures of relapses or progression of disability, these 3 cohorts seem to behave very similarly and comparably. On the other hand, when we add in also a measure of the effect on gadolinium enhancing lesions, what we see is that interferon actually has a much superior impact on contrast enhancing lesions.

23. Surrogate endpoints may not be universal (Fig. 3)

So going back to our model, what this suggests is that interferon does in fact act on a causal pathway that leads to formation of contrast enhancing lesions and most of its effect on contrast enhancing lesions can explain its effect on suppression of

Fig. 3　Surrogate endpoints may not be universal



relapses. Copaxone® on the other hand seems to have less of an effect on suppressing contrast enhancing lesions, and yet it suppresses relapses just as effectively. This suggests that Copaxone® is probably also acting on another causal pathway which independently suppresses relapses. So if we go just by that surrogate endpoint, we would say that interferon is much more effective than Copaxone®, and we would be wrong.

24. Identification of a novel biomarker

The daclizumab study was important also for another reason. It also led to the identification of a new novel biomarker. In fact, it was seen that in patients treated with daclizumab, there was an expansion of a subset of the innate immune system called CD56 bright natural killer cells. The expansion of this population was able to differentiate between full and partial responders essentially. So the patients who had a dramatic expansion in this subpopulation were in fact the patients who were full responders compared to the patients who were partial responders.

25. Expansion of CD56 ^{bright} NK cells correlates with decrease in CEL

It would seem that expansion of this population correlated very well with the decrease in contrast enhancing lesions. This seems to be in fact a very good biomarker for predicting response to therapy, and can certainly be obtained much earlier than measurement of contrast enhancing lesions or even relapses, which would take even longer.

26. IT/IV rituximab for SPMS (RiVITALISe)

I'd like to tell you about another trial that is ongoing right now at the NIH. It's called RiVITALISe, and this is a trial of treating secondary progressive multiple sclerosis with intravenous and intrathecal rituximab. The goal of this study is to try to eradicate the presence of ectopic lymphoid follicles which are present in the meninges of patients with secondary progressive MS. They're thought to be on the causal pathway leading to disability in this particular variant of MS. Rituximab obviously targets the B-cells, and we would like to eliminate those B-cells using this strategy. Rituximab has never been administered to secondary progressive patients in this way; so the calculations for the dosing were based on other medical conditions such as cancer or other autoimmune diseases.

27. RiVITALISe

We were not sure if we were adequately dosing our patients to achieve our goal. For this reason, we have included in our design two interim analyses. The first interim analysis was performed after the first 3 patients received active drug and this was done to assess the level of B-cell depletion. What we found was that only one of these three patients actually demonstrated any B-cell depletion. So we were allowed by protocol (as it was pre-planned) to modify the dosing of our Rituximab regimen. Now we have a second interim analysis which is a futility analysis, and this has not yet taken place but will take place after 11 patients have been dosed with active drug. In this analysis what we will be doing is measuring B-cell depletion as well as some cytokine markers of B-cell func-

tion, and we have specific stopping criteria. If we don't see at least a 50% decrease of B-cells as well as related changes in the cytokine measurements, the trial will be stopped for lack of biological effect, and therefore, maximizing resources and not continuing a trial that is not likely to succeed.

28. Model of primary progressive MS

I'd like to share with you another clinical trial that is ongoing at NIH right now. This is for Primary Progressive MS, another variant of MS. This variant of MS does not have a great degree of inflammation. It is felt that the disability which develops in these types of patients is very gradually and very slowly manifested, and is most likely a product of neurodegenerative process of which the causal pathways are simply not known. So we don't have any surrogate endpoints that we can measure. The only measurements that we have are very poor and take a very long time to develop because it is a slowly evolving disease.

29. IPPoMs (Fig. 4)

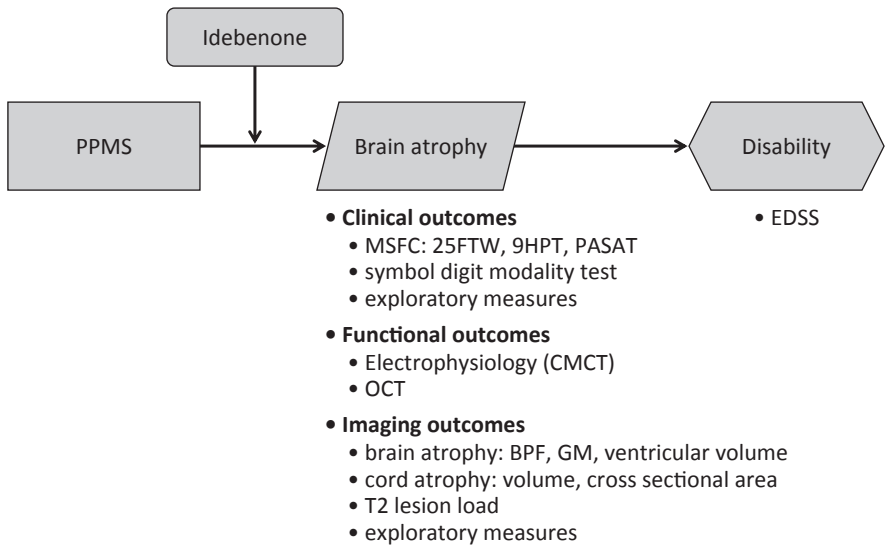
So in this disease, what we are doing is, we are administering a drug based on evidence that mitochondrial dysfunction might be playing a role in the neurodegenerative process. The drug is called idebenone, and it does improve mitochondrial function.

We have two objectives in this study. The first one is to test the mechanistic hypothesis that mitochondrial dysfunction does indeed play a role in this neurodegenerative process in primary progressive MS. The second is that we wish to identify the most promising outcome measure or measures in this disease, so that for future trials we may be able to rapidly screen other treatments.

So the study is powered on datasets of brain atrophy in primary progressive MS. However, we are using an adaptive trial design in which we will be able to select or change our endpoint based on the initial results.

The way this is working is that brain atrophy is

Fig. 4 IPPoMS



the surrogate endpoint, the primary surrogate endpoint, but we are collecting a large number of clinical outcomes, functional outcomes as well as imaging outcomes.

All of these measures are being collected during a one year pre-treatment baseline on all the patients who are enrolled in the study.

There is an interim analysis prospectively planned, and which will be conducted once 30 patients have concluded this one year of data collection. At that point what we will assess is which of all these measures is the most sensitive to change over a period of one year. We will then select that measure as the primary outcome for the trial in the hopes that this will allow us to measure treatment effect more effectively.

30. Surrogate endpoint: Gray matter atrophy

The 30 patients have actually completed this beginning phase of the study and the analysis is underway. Unfortunately, I don't have those results to share with you. But I would like to show you some very preliminary data which are based just on, I think, 18 patients. The first thing is the measurement of brain atrophy is not trivial and we felt that we may be able to improve the methods of measuring brain atrophy. One of the things we thought was that gray matter atrophy might be more sensitive because it is the gray matter which is mostly affected in primary progressive MS. What we found unfortunately is that gray matter atrophy is not a very reliable measure. There's a lot of variability, and so we don't feel that we will be able to demonstrate much of a treatment effect with this measure.

31. Surrogate endpoint: Ventricular volume

A second measure is ventricular volume as a measure of brain atrophy. This measure seems to have improved characteristics with much less variability. So this may end up being the imaging measure with the best characteristics.

32. Surrogate endpoint: 25 Foot Timed Walk

Another measure that we are using is the 25 Foot Timed Walk. Unfortunately in most of our patients, 25 Foot Timed Walk does not appear to be sensitive enough to change over this brief period of one year. Even taking out the outliers, we still see that most patients don't have really significant change over this one year of time. I will note that 25 Foot Timed Walk was the primary outcome measure and led to the regulatory approval of dalfampridine for symptomatic management of relapsing remitting MS which just goes to show again one surrogate endpoint for one study, one patient population, cannot always be automatically applied to another.

Another feature of the 25 Foot Timed Walk which is disturbing to us is the fact that the variability of this measure depends on how disabled the patient is. So that the more disabled the patient is, the less reliable this measure tends to be. And Primary Progressive MS patients tend to be quite impaired or quite disabled.

As one could imagine then, if we plot the 25 Foot Timed Walk, we see that there is definitely a eschewed distribution.

33. Surrogate endpoint: 25 Foot Timed Walk vs. Foot Tapping Speed

We have an alternative, an exploratory clinical measure, which is foot tapping speed. When we plot the foot tapping speed, we actually get a pretty normal distribution.

When we perform Bland-Altman analysis, we see that foot tapping speed appears to have constant reliability across the range of disability, while 25 Foot Timed Walk as already mentioned does not. We don't know yet if Foot Tapping Speed is going to be sensitive enough to change over a short period of time such as one year. But it certainly does have some features that make it an attractive measure to explore further in this disease.

34. NIB Optic Neuritis Study (Fig. 5)

I want to tell you about one more study which will be starting up shortly. This is an Optic Neuritis Study of the Neuroimmunology Branch (NIB). The goal is to identify a surrogate endpoint that can be used in a future trial of neuroprotection or neurorepair.

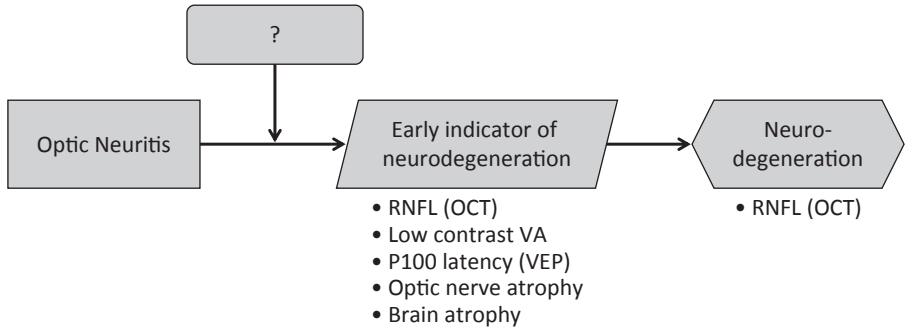
Following an episode of optic neuritis, most patients will recover usually within a period of about a month – either with full return of function or with some impairment in function. But what happens after that is a slow neurodegenerative process of the optic nerve which can be measured over the course of one year by ocular coherence tomography, and specifically, with the reduction in the thickness of the retinal nerve fiber layer.

What we would like to do is to be able to measure this neurodegenerative process early to find an early indicator of neurodegeneration that can predict this end neurodegeneration. The problem is that the measures of structure and function that we have right now simply are not sensitive enough to be able to detect the change in this early phase over a short period of time. So again, our goal is to be able to detect. If this is when the optic neuritis occurs (shown on the slide), to detect the change in the few months immediately following recovery from the optic neuritis, that can predict a more distal endpoint.

35. Composite measure of structure & function of the visual pathway

So we will try to develop a composite measure which integrates the information from ocular coher-

Fig. 5　NIB Optic Neuritis Study



ence tomography, low vision contrast acuity, functional studies of visually evoked potentials, optic never atrophy as well as brain atrophy. Our hope is that we may be able to take advantage of this composite measure, borrow strength from multiple outcome measures, increase the power and have increased ability to detect a treatment effect, maybe shortening trial and decreasing sample size. In addition we have built-in consistency check because we can look to see that all these measures are going in the right direction together.

Disadvantages of composite measures do exist. One of them is that we may dilute our treatment effect if one of our outcomes is not responsive to change or goes the wrong direction. It obviously makes the study more difficult, requiring more tasks for the patient, so increased costs. It can be difficult to generate a score that is easily interpretable from composites, so this might lead to reduced acceptance of this measure such as we have seen initially with MSFC in multiple sclerosis. But we are hopeful that this strategy will help us to be able to develop a model that might be used for future trials of neuroprotection.

36. Concluding thoughts

I just want to share some concluding thoughts. These are actually the words of Janet Woodcock who is the director for the Center for Drug Evaluation and Research at the FDA. She's a strong supporter of the use of biomarkers in drug development as well as surrogate endpoints. What she has said is that biomarkers must be used in order to be accepted. That means they have to be incorporated into clinical trials. Obviously, the add-on costs to clinical trials have been a barrier to incorporating biomarker development into trials. But what she has said and what she supports is that we must have collaboration between government, academia and industry in order to promote the development of biomarkers and surrogate endpoints.

So with that, I would like to thank all of my colleagues back at home, and thank you for listening.

*　　*　　*