Biomarker Development, Methodological Challenges

Gary R. Cutter, PhD

Emeritus Professor of Biostatistics
Department of Biostatistics
University of Alabama at Birmingham School of Public Health

ABSTRACT

Biomarker development is a common endeavor in medical research. The purpose is to find indicators of disease occurrence or prognostic markers for response. The process of development of biomarkers often starts with showing mean differences between responders and non-responders or those with a disease or condition versus those without. However, these statistically significant mean differences, while necessary are not sufficient to validate a biomarker. Sensitivity, specificity, positive and negative predictive value are at least as important and the relative increase in performance using the biomarker over the usual clinical variables should be demonstrated. This paper discusses the various assessments in the context of use for the biomarker, the need for characteristics in addition to mean differences and the importance of independent validation of putative biomarkers. Lastly, it is hoped that the process and thoroughness necessary be considered with recognition that the task is at best difficult.

Key Words: Biomarkers, Surrogate Outcomes, Prentice Criteria, Prognostic Biomarkers, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, Validity

Introduction

The search for biomarkers is not new. Fever has long been used as a sentinel biomarker for illness in the body. This common tool for lay and professionals alike is, of course, a consequence of disease rather than a predictor of disease, although it may be a harbinger of a consequence indicative of the need for treatment or the impending consequences of disease. Often in the search of biomarkers we use a similar fallacy called the *post hoc ergo propter hoc* fallacy, whereby one assumes that one event must have caused a later event simply because it happened after the other. One might argue that this happens with acetylcholine receptor (AChR) antibodies in myasthenia gravis. The fact that these are defining the disease does not mean that the severity or course of disease is predicted or identified by the levels seen. Prior occurrence is not sufficient to define a predictive biomarker. Biomarkers may indicate what will happen or they can be useful to avert something happening.

"Biomarkers are biological substances, characteristics, or images that provide an indication of the biological state of an organism." (group 2001) (Medicine 2009). The FDA defines 5 categories that need to be considered when developing or evaluating a biomarker:

- Context of use (purpose, population, and nature of disease)
- Analytical validity
- Clinical validity
- Clinical utility
- Gold standard validation

The above categories are somewhat self-evident. The context of use (FDA) or COU in FDA nomenclature, defines two steps in the development of a biomarker. First is the category of use into one of 7 categories: Diagnostic; Monitoring; Predictive; Prognostic; Pharmacodynamic/Response; Safety; Susceptibility/Risk. Then within each category, there is the determination of how the biomarker will be used. For example, the diagnostic use might be for subject selection in a trial: an AChR antibody test might be the cardinal biomarker of myasthenia gravis (MG) and the level might be used to quantify the selection criteria for qualification for a trial as was done in the Thymectomy Trial in Non-Thymomatous Myasthenia Gravis Patients Receiving Prednisone Therapy (MGTX) trial (Wolfe GI 2016) .

Often biomarkers are classified in other related ways, such as a *surrogate* endpoint which is assessed pre- and post-treatment as an early measure of clinical outcome; a *pharmaco-dynamic* biomarker which is assessed pre- and post-treatment as a measure of the effect of treatment on disease; a *prognostic* biomarker, to identify which patients need treatment; and a *predictive* biomarker to determine which patients are likely to benefit or respond from a specific treatment.

Biomarkers aimed at treatment should be able to improve on the prediction of responders over the clinical variables available. That is, having the biomarker results in hand should lead to better prediction of the likelihood of response. Thus, biomarkers may improve treatment decisions by identifying responders in general or identifying treatments that work better in subgroups or vice versa. One example might be the muscle-specific receptor tyrosine kinase (**MuSK**) which identifies patients who are less likely to respond to conventional MG treatments. There are a number of ways statistically that this can be done: Show that the area under the receiver operator characteristic (ROC) curve is increased (Pencina MJ 2010); achieve improvements in the net reclassification index (Hlatky MA 2009); use the integrated discrimination index (IDI) (Pencina MJ 2008). Each of these measures are calculations that return a number that is used to assess if the classifications have been improved by the addition of the biomarker to the prediction equation. The increase in the area under the ROC curve is commonly used, indicating improved sensitivity and/or specificity of the

biomarker under consideration, but is only an indicator of improvement and does not always imply the improvement is clinically meaningful. Thus, a combination of statistical tools is needed to assess the added value of the biomarker.

Surrogate Biomarkers

Validating a biomarker as a surrogate for a clinical outcome is extremely difficult. Usually this requires a series of randomized trials with both the biomarker and clinical outcome measured demonstrating correlated differences in the outcome and/or mediation of the treatment effect by the biomarker. While there are criteria for defining when this has occurred, it is rare that such surrogates can be found. Even the concept of surrogate is dubious because often a large treatment effect on the surrogate corresponds to only a small treatment effect on the true clinical outcome. Think of blood pressure treatment for hypertension and the outcome of cardiovascular disease. Blood pressure treatments often lower blood pressure by 15% to 20%, but the impact on mortality may be less than 5%. However, on a population level this impact is large and clinically meaningful indicating again the context of use is important.

Prentice (RL 1989) created what may be considered the most stringent criteria or the goal of a surrogate outcome. Within a randomized clinical trial (RCT):

- The treatment must have an effect on the surrogate.
- The treatment must have an effect on the clinical outcome.
- The surrogate and the clinical outcome must be correlated.
- The treatment effect on the true clinical outcome must disappear after adjusting for the surrogate.

The last of these criteria is, for the most part, unachievable. It is this last criterion that is often relaxed to significantly mediate the outcome and link the concept of a surrogate to a mediating variable. Thus, a surrogate endpoint (Biomarker) is said to be an intermediate (instrumental) variable that can be used to indicate the true clinical endpoint. If the full effect of treatment on the responder status is mediated through the biomarker, then we have a surrogate as defined by the Prentice criteria.

Prognostic Biomarkers

Most prognostic factors are not used, because they are not therapeutically relevant. For example, age is strong predictor of poor outcomes in many situations, yet it is not something we can intervene on therapeutically. We want prognostic biomarkers in the concept of surrogates, which are subject to manipulation and therapeutic intervention. However, to develop such markers requires carefully designed studies even though many are identified via retrospective analyses of existing datasets. That said, most prognostic factor studies are poorly designed. They are not focused on a clear therapeutic decision context and often use a convenience sample of patients for whom material or information is available. Generally, the patients are too heterogeneous to support therapeutically relevant conclusions, and, commonly, they address statistical significance, rather than predictive accuracy, relative to standard prognostic factors.

Two examples might help clarify these issues. Low density lipoprotein receptor-related protein 4 (LRP4-Ab) has recently been considered as a potential biomarker in seronegative MG patients (Chung HY 2023). These authors attempt to develop a cell-based assay (CBA) for the detection, however, they report that "there is no gold-standard test for LRP4-Ab that can be used to compare the performance of the present CBA. The possibility of false-positive results cannot be ruled out. Further studies using different methods for detecting LRP4-Ab are necessary." This lack of a gold standard for validation is equally important with clinical outcomes, which often use a specified amount of change, such as 2 or 3 points on the MG-ADL scale as indicative of being a responder. This often ignores the recruitment requirement to have scores above some cut point, such that responders are mixed with individuals measured in error at baseline with values higher than they actually are. This leads to regression toward the mean and in a randomized trial is expected to be the same in both treatment groups, but in biomarker discovery is confounded with response. Another example is the use of statistically significant differences to infer biomarker status. In the paper by Cavalcante et al. (2019), a microRNA signature was associated with being a biomarker of responsiveness to treatment in MG, and while significant differences are seen, the sensitivity is only around 50%.

Predictive Classifiers

Many treatments benefit only a minority of patients to whom they are administered. This is particularly true for molecularly targeted drugs. Predictive classifiers seek to be able to predict which patients are likely to benefit and which patients can be saved from unnecessary toxicity. Thus, predictive classifiers are focused on the benefit/risk equation of treatment and enhance the patient's chance of receiving a drug that helps them or does not hurt them. If we knew that a person/patient with a specific HLA type when given a certain drug has a higher likelihood of drug-induced liver injury, we might avoid the use of this treatment in favor of some other. Similarly, if we know that a specific HLA type responds better, we would use the treatment associated with the better response. These biomarkers can help control medical costs while improving the success rate of treatment and even clinical drug development.

Validity

Validity implies correctness, but it requires more than simply opinion or face validity. It should demonstrate that the biomarker is predictive *a priori* rather than *a posteriori*. Even though identification and performance characteristics

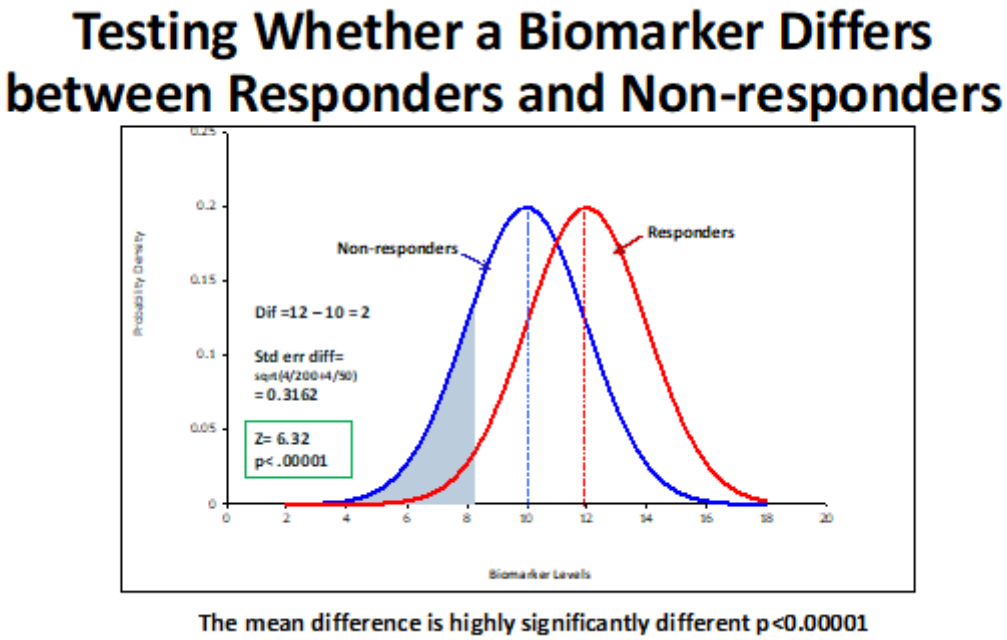

Figure 1: Relative Frequency of Biomarker Levels in Responders (red) and Non-Responders (blue)

are often evaluated by comparing cases to controls, the true test is from prospectively applying the putative biomarker in studies or trials that demonstrate the predictive value. Consider a biomarker for disease diagnosis. Was there an independent, blind comparison with a reference standard of diagnosis? Was the test evaluated in an appropriate spectrum of patients (like those actually seen in clinical practice, where there is diagnostic uncertainty)? Was the reference standard applied regardless of the diagnostic test result? When tests are invasive or expensive, we often only perform these after a higher suspicion of disease is present, this leads to verification bias. For example, because of cost, yield and small risk, routine CTs as the gold standard for detecting thymomas are given to patients only when symptoms are present. Thus, a study of a biomarker for thymoma might underestimate false negatives because patients with symptoms under the threshold were not offered a CT. Additionally, for establishing a biomarker, it is important to ask whether the test is validated in a second group of patients. The last of these questions is essential to provide independent confirmation of the value of the biomarker.

As noted above, too often developers of biomarkers use statistical significance of differences between those with the disease compared to those without the disease as evidence for a putative biomarker. Let's look at an example. Suppose we want to assess whether a biomarker differs between responders and non-responders.

Suppose amongst **non-responders** to CMP (Cutter's Magic Potion) the mean interleukin-17 (IL-17) was found to be 10 with a standard deviation of 2 (sample size of n_1=200). In **responders** it was found, on average, to be 12 with a standard deviation of 2 (sample size of n_2=50). Is IL-17 a biomarker of response? Figure 1 shows the hypothetical distribution of IL-17 for responders (red frequency distribution) and non-responders (blue frequency distribution). The blue curve to the left shows the distribution of the non-responders and the red one to the right are the responders. Approximately 10% of the responders had levels of IL=17 lower than a little over 8 (shown as the shaded area on the blue non-responders curve).

As is often done by researchers when they are attempting to identify a biomarker, they will test the mean differences between responders and non-responders or cases versus controls to convince the reader that the biomarker is indeed a predictor of response. Here we see a mean difference of 2 units (mean of 12 for responders and 10 for non-responders). The t-test for the difference uses the standard error of the mean difference between responders and non-responders to decide if this difference is larger than that expected by change, and this takes into account the standard deviation of the responders and non-responders and the respective sample sizes.

Thus, standard error of mean difference is:

= square root of (variance in non-responders/n_1 + variance in responders/n_2)

= sqrt (4/200+4/50) = 0.3162

And the t-test for the difference between the two groups:

= 2/0.3162 = 6.33 yielding a p-value of 0.00001

This tells us that the means are significantly different, but is this sufficient to establish IL-17 as a biomarker of response? Many investigators think this is so, but while this result is necessary, it is not sufficient. There are other summarizations that are important and meaningful. Four of them are: Sensitivity, which is the probability of a positive test among patients with disease; Specificity, which is the probability of a negative test among patients without disease; Positive Predictive Value (PPV) and Negative Predictive Value, (NPV). PPV means of those that have a positive test, the probability that the individual has the disease or condition (or doesn't have the disease or condition – NPV). The former two, sensitivity and specificity, are what developers of biomarkers generally focus on; however, PPV and NPV are the most important to the patient. Why? While sensitivity, specificity, and false positives and negatives help a discipline, the clinician or patient decide whether to advocate for a biomarker being useful or perform a test with a biomarker because it is useful; patients (and their clinicians) are not directly interested in false positives and false negatives, once they have the result. They want to know what the test means for them! "I have a positive test – what does that mean for me?" For example, if the sensitivity of mammography for detecting breast cancer is 75% and specificity is 98%, this may help policy makers and clinicians recommend a mammogram. However, because so many more women do not have cancer, the false positives greatly outnumber the true positives with this screening test (biomarker). Thus, a clinician can be a calming force for a woman with a positive mammogram informing her that of those with a positive mammogram only about 10% actually have breast cancer. The clinician is using the positive predictive value to assuage the panic of the positive mammogram.

Let's look a bit closer at sensitivity and specificity in our IL-17 example from Figure 1. Recall from Figure 1, that the mean IL-17 in responders was 12 and in non-responders it was 10. If we use 10 as our critical value for determining sensitivity and specificity for those above and below the mean of the non-responders, we would ask in assessing if IL-17 is a biomarker for response, what is the probability of being a responder if their IL-17 is above 10? Similarly, what is the probability of being a non-responder if their IL-17 is below 10. In Figure 1, we see that for responders 10 is 1 standard deviation below the mean (recall the standard deviation is 2 and thus 1 standard deviation below the mean of 12). This translates into 64% of the responders being above 10 (this results from assuming a normal distribution of the IL-17, where 1 standard deviation below the mean separates the population into 64% above the -1 standard deviation and below -1 standard deviation). Similarly, among non-responders, the mean was 10 and thus 50% of the non-responders are below 10 (in a normal distribution 50% are below the mean). If one used IL-17 as a biomarker with the value set at 10, it would not be a good biomarker because so many participants would be misclassified: 50% of the non-responders would be above 10 and thus false positives! In the responder predicted category, 36% of the responders would be below 10 and thus false negatives.

What were the PPV and NPV from Figure 1? There were 200 non-responders and 50% of them are expected to be above 10 or 100 individuals. Of the 50 responders, 64% or 32 were above 10. Thus the PPV = 32/(100+32) = 0.242. Stated another way, if your IL-17 was above 10, you had a 24.2% chance of being a responder. If you just had historical data and no putative biomarker, you would guess that 50/(200+50) = 20% would be responders. This naïve estimate (not taking into account the biomarker) is only slightly below the information that is coming from having the biomarker, that is 24.2% compared to 20%. Thus, while it is an increase in the estimated chance of response, it probably is insufficient to convince users that it is a relevant biomarker.

It is also important to remember that positive and negative predictive value depend on the prevalence of the disease or the outcome. Myasthenia Gravis is estimated at a prevalence of 20 per 100,000 population. Suppose we develop a questionnaire that we think can identify MG. In the clinic we show it has 95% sensitivity in correctly identifying the MG patients, but only 90% specificity, what is the positive predictive value? Consider 100,000 individuals evaluated in a population survey. We expect 20 cases with a sensitivity of 95% and thus, 19 of the 20 cases would be positive on our questionnaire. However, because the specificity is only 90% of the 99,980 individuals without MG, 10% or approximately 10,000 would be flagged as potential cases. Our PPV would then be 19/10,000 or 0.19% and virtually an NPV of 1.

Another common approach to establishing a biomarker is to compare the extremes of the distribution of the biomarker. Investigators often compare the lowest decile or quartile to the highest decile or quartile to show their biomarker works. This too is necessary for a biomarker's performance, but it is not sufficient to establish a biomarker. Consider the lower quartile compared to the upper quartile. Increased response in one quartile compared to the other still leave 50% (quartiles 2 and 3) out of the quantification. This can lead to substantial misclassification and poor performance by the biomarker. The value as a biomarker actually then relies on what happens in the middle rather than at the extremes. There is an especially prevalent use of these extreme comparisons in epidemiological studies and specifically diet studies. Part of the rationale for this prevalent use is that diet is poorly measured and thus the misclassification is not from the performance of the biomarker, but rather the error in assessment of the underlying diet. This may be true, but one needs to exercise caution when interpreting a biomarker determined solely on the basis of comparison of the extremes. In the search for biomarkers, statistically significant differences between

these groups are necessary BUT NOT SUFFICIENT. Achieving high levels of sensitivity and specificity require low variability within a population and high variability between populations and good biomarkers or classifiers require high sensitivity, specificity, PPV and NPV.

The Sequential Organ Failure Assessment or SOFA score is a widely used biomarker of disease prognosis. It has been shown to predict mortality in a variety of settings from the intensive care unit to results of COVID-19 infection. AChR and Aquaporin4 are often thought of as biomarkers, but since they are often used in the definition of the disease and do not clearly associate the levels found with prognosis, they fail to meet these requirements. CD4 counts in HIV and/or hemoglobin A1C in diabetes have been successfully used to characterize these as biomarkers. Although they fail to meet the Prentice criteria cited above, they have proven to be very important biomarkers of response.

Quite common in the development of biomarkers, is the question: how many or much more do I need? This question often comes to biostatisticians brought in to help "bless" a biomarker being considered. While this is a reasonable question, especially in this era of adaptive designs, where incrementally evaluating data is used to arrive at a more firm conclusion, it is also a problematic question. This is because the biostatistician doesn't know what has been done to get to this point in the research. Were outliers tossed, samples rerun, was the development of the data done under a defined or strict protocol or has this evolved and the researcher gained interest in the putative biomarker with further experiments and analyses? While it is important and natural to conduct exploratory data analyses to develop a biomarker, the process is not a continuous one. At some point in the development, a more formal evaluation should occur. This often is done by adding the formal evaluation to a clinical trial providing objective and rigorous evaluation of the putative biomarker. Irrespective of whether this is done within a trial, a formal evaluation under a defined protocol is essential. Adaptive designs require carefully crafted protocols to ensure adequate control of type I errors and *a priori* decision-making.

We are in the era of digital and remote monitoring which will lead to more and more putative biomarkers. The digital biomarker development process has been categorized (Bent B 2020) into: State the goal; define the sensor data to be used; specify other data needed; define the preprocessing necessary; perform exploratory data analyses to evaluate relationships; identify feature engineering and feature selection. What seems missing from this development process is the utility of the biomarker or biosensor. Defining the context of use and the utility in that context are often ignored as the rush to apply or market the device occurs. The utility is often assumed or implied, but not formally evaluated. This last step is critically important lest the information derived from the device is of limited value clinically.

Some digital biomarkers have been shown to improve care. Digital glucose monitors which free the patients from finger sticks and provide real time monitoring of blood glucose continue the known benefits of tight control in diabetes. The plethora of step counters, however, have not been shown to provide improved health despite their widespread use other than in small studies and anecdotal experiences. This latter example exemplifies several issues. First is the rapid escalation in the availability of digital monitoring and the benefits may take much longer to assess. Studies of the control of mild hypertension and tight control of diabetes evaluated mortality over a 5-year period and of course took several years longer in real time to get answers due to funding, initiation, recruitment, etc. In addition, there are the concepts of efficacy and effectiveness. Can the digital monitor work as proposed, that is, is it fit for purpose. These are issues with home step counters and home pulse oximeters. Then, assuming they achieve the technical details of measuring what they purport to measure, do they, in ideal settings, change the clinical outcome (efficacy)? Finally, if they work and possess efficacy, do people use them? The use in practice results in effectiveness and incorporates both accuracy and precision of the device with efficacy and individual compliance.

On the other hand, even small increments in some biomarkers can be important. If we can develop behavioral threat assessments for mass shootings as biomarkers and they lead to actions and/or interventions that prevent mass gun violence, then the biomarker doesn't have to have great sensitivity to be valuable. As long as there are few negative consequences for the false positives, even a poorly performing biomarker might be helpful. The benefit is great, and risk is low or non-existent. Thus, the question being addressed is central to the interpretation of the purported biomarker.

A final word of caution. Developing biomarkers is harder than most investigators think. Without validation, and independent validation, they are just another outcome measure. Investigators need to remember the difference between a correlate and a surrogate. Further, while the excitement of finding mean differences on a putative biomarker are encouraging, mean differences are necessary but not sufficient to establish a biomarker.

References

Bent B, W. K., Grzesiak E, Jiang C, Qi Y, Jiang Y, Cho P, Zingler K,Ogbeide FI, Zhao A, Runge R, Sim I, and Dunn J (2020). "The digital biomarker discovery pipeline: An open-source software platform for the development of digital biomarkers using mHealth and wearables data." Journal of Clinical and Translational Science **5**: 1–8.

Cavalcante P, M. T., Barzago C, Scandiffio L, Bortone F, Bonanno S, Frangiamore R, Mantegazza R, Bernasconi P, Brenner T, Vaknin-Dembinsky A, Antozzi C. (2019). "MicroRNA signature associated with treatment response

in myasthenia gravis: A further step towards precision medicine." Pharmacol Res(Oct; 148:104388): 1-13.

Chung HY, K. M., Kim SW, Oh J, Shin HY (2023). "Development and Application of a Cell-Based Assay for LRP4 Antibody Associated With Myasthenia Gravis. ." J Clin Neurol. **19**(1): 60-66.

FDA, C. o. U. group, B. d. w. (2001). "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework." Clinical Pharmacology and Therapeutics **69**(3): 89-95.

Hlatky MA, G. P., Arnett DK, Ballantyne CM, Criqui MH, Elkind MSV, Go AS, Harrell FE,Hong Y, Howard BV, Howard VJ, Hsue PY, Kramer CM, McConnell JP, Normand SL, O'Donnell CJ, Smith SC, Wilson PWF (2009). "Criteria for Evaluation of Novel Markers of Cardiovascular Risk." Circulation **119**(17): 2408–2416.

Medicine, I. o. (2009). "Accelerating the Development of Biomarkers for Drug Safety: Workshop Summary."

Pencina MJ, D. A. R., D' Agostino RB, Vasan RS (2008). "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. ." Statistics in Medicine **27**(2): 157–172.

Pencina MJ, D. A. R., Vasan RS (2010). "Statistical methods for assessment of added usefulness of new biomarkers." Clin Chem Lab Med **48**(12): 1703–1711.

RL, P. (1989). "Surrogate endpoints in clinical trials: definition and operational criteria." Statistics in Medicine **8**(4): 431-440.

Wolfe GI, K. H., Aban IB, Minisman G, Kuo HC, Marx A, Ströbel P, Mazia C, Oger J, Cea JG, Heckmann JM, Evoli A, Nix W, Ciafaloni E, Antonini G, Witoonpanich R, King JO, Beydoun SR, Chalk CH, Barboi AC, Amato AA, Shaibani AI, Katirji B, Lecky BR, Buckley C, Vincent A, Dias-Tosta E, Yoshikawa H, Waddington-Cruz M, Pulley MT, Rivner MH, Kostera-Pruszczyk A, Pascuzzi RM, Jackson CE, Garcia Ramos GS, Verschuuren JJ, Massey JM, Kissel JT, Werneck LC, Benatar M, Barohn RJ, Tandan R, Mozaffar T, Conwit R, Odenkirchen J, Sonett JR, Jaretzki A 3rd, Newsom-Davis J, Cutter GR. MGTX Study Group. (2016). "Randomized Trial of Thymectomy in Myasthenia Gravis." N Engl J Med(Aug 11;375(6)): 511-522.