[Going beyond design, going beyond intervention: The American Academy of Neurology Clinical Practice Guideline Process and other approaches.](https://www.ktdrr.org/training/webcasts/webcast10-13/11/index.html)

**Presenter:**

**Dr. Marcel Dijkers**

A webinar aired June 18, 2014, sponsored by SEDL’s Center on

Knowledge Translation for Disability and Rehabilitation Research (KTDRR)

Funded by NIDRR, US Department of Education, PR#H133A120012

Edited transcript for audio/video file on YouTube: <http://youtu.be/V8Pyf91ZsgM>

 >> ANN WILLIAMS: Thank you all for joining us. I am Ann Williams of KTDRR at SEDL in Austin, Texas. And I will be moderating today's webinar entitled Going beyond design, going beyond intervention: The American Academy of Neurology Clinical Practice Guideline Process and other approaches.

This is the second in a series of four webinars focusing on systematic reviews, from evidence to recommendations.

Before we begin I'd like to go through some of the Adobe Connect logistics. You all should be listening to this presentation through your computer speakers. If you need to turn up the volume, you can do so on your own computer in your audio settings. If you have any questions or comments, please type them in to the chat box on the left side of the screen and Marcel or I will address these as appropriate. We have CART captioning that's available. And the link to the CART is in the useful links pod on the right‑hand side of the screen. And it will open a link to another window when you click on it.

Finally a pdf and a text version of the powerpoint are available in the pod labeled Useful Documents. So if you would like to download a copy, please feel free to do so.

I would like to thank my colleague Steven Boydston for his logistical and technical support for today's session. This webinar series is offered through the Center on Knowledge Translation for Disability and Rehabilitation Research or KTDRR, which is funded by the National Institute on Disability, Independent Living, and Rehabilitation Research. KTDRR is sponsoring a community of practice on evidence for disability and rehabilitation research, and this series of webinars addresses systematic reviews with a special focus on what is considered evidence and why and how this evidence is qualified, synthesized and turned in to recommendation for clinicians and other practitioners.

Now it is my pleasure to introduce Marcel Dijkers from the Icahn School of Medicine at Mount Sinai. Dr. Dijkers is a research professor in the Department of Rehabilitation Medicine and a senior investigator in the Brain Injury Research Center. He is also the director of the NIDRR‑funded disability and rehabilitation research project on classification and management of medical rehabilitation interventions as well as the Mount Sinai advanced rehabilitation research training project. He is the senior investigator for the New York TBI model system funded by NIDRR. Dr. Dijkers has published more than 120 articles and chapters on disability and rehabilitation research. So without further ado, Marcel, are you ready?

 >> MARCEL DIJKERS: I am ready. Thank you Ann and we have Ann today as our hostess because Joann Starks who normally takes care of these duties is on the road, out of the country in Toronto. We are very glad to have Ann. Our objective is unchanged from what you saw previously. We want to get a conversation going about, you know, what do we call evidence, why and how do we massage the research results to have evidence and how does it from there become recommendations.
 So today is the second in a series of four and I will skip through this. The last time at our first session we talked about the role of research, specifically clinical research and decision making of practitioners and with quite some discussion of studies and how those might have been turned in to evidence‑based practice resources.
 We went over the process of creating a systematic review, how it is done and what are the steps. We talked about the meaning of evidence and the need to evaluate the quality of evidence resulting from clinical research, you know, talking about what generally people call design and which I like to call design with a big D. Is it randomized clinical trials or is it pre/post study versus a single study design, et cetera. The quality of evidence is determined by that but also what goes on with what I like to call little d may be referring to "details" of implementing, designing and implementing research, the like subjects being blinded or what's the percent of follow‑up that's being achieved, et cetera.
 Then we talked about the hierarchies for evidence relevant to interventions developed by Sackett which is already 30 years old and more or less is the whipping boy. And we also looked at Cicerone which is close to 15 years old. And if we would ask Mr. Cicerone, he might wish that he could start from scratch and do it all over with more sophisticated approach.
 We talked about checklist and rating scales for evidence quality such as the Jadad scale and the PEDro scale, beloved by psychotherapists or physical therapists as we call them this side of the border. And we discussed possible uses of checklists and rating scales.
 So if you have any questions about, you know, what I have spoken about this far, up to and including what was presented the last time, please don't hesitate. And in the meantimeI will go forward to the next slides. There is more to evidence‑based practice than randomized control trials and a lot of these slides will just have the abbreviations. So try to keep in mind to what those three letter abbreviations mean. Of course RCTs are relevant to if we want to collect information, specify evidence on prognosis, diagnosis, screening, et cetera, et cetera.
 And we may need to other research designs are relevant. We may need different hierarchies for the other purposes for which we do systematic reviewing. We may have multiple hierarchies, one each for specific research question type.
 And indeed those have been developed since Sackett published his simple hierarchy for intervention studies. One is produced by the Oxford Centre for Evidence‑Based Medicine and they have quick suggestion ‑‑ succession presented two iterations of what they propose and we'll go a little bit later back to the 2009 hierarchy but first let's talk about the 2011 hierarchy of evidence. So they distinguish six different questions that people who are in to evidence‑based medicine from our perspective we might as well say just the same evidence‑based practice and other disciplines. Questions that people may have. How common is the problem? Is this diagnosis or the monitoring test that we are using accurate. Questions of diagnosis. What will happen if we don't read this patient's prognosis? Does this intervention actually help? What and how strong are the treatment benefits and what are the common harms that we can expect from this treatment? And then, you know, what are the more rare harms? And lastly if we are talking about screening for particular conditions is it worthwhile to do this screening? And even might think about just things as breast cancer screening. And you are probably aware of the big fight that have existed and still exist whether it is worthwhile to do that.
 So if you reference this to what Sackett did, the brown column, does this intervention help? Is the only one of this hierarchy created. So the Oxford people have to provide five more.
 Before we get in to some details, some notes the scheme that you can find on the Oxford website is optimized not for people who create systematic review but for the people who need to use evidence, especially people who need to do that quick, about to call that bedside evidence‑based practice where you discover a particular problem in one of your patients and you want to ‑‑ if so what's the best way of treating that. Let's see what does the evidence say. Now generally you as in practitioner would be happiest with finding a systematic review that covers your question. For the Oxford people systematic review is always the highest rank in the hierarchies. Also a very high one is one that is unlikely to ever occur which is an N-of 1 trial that was conducted with the patient that the clinician wants to treat. So essentially end of one trial it is terminology that comes out of medicine but it is a single subject design with a very long ABABAB sequence randomized, et cetera.
 So yes, if you want to treat a patient with a new problem you grab for the literature, the No. 1 would be the systematic reviews. You probably if you already did an N-of 1 trial, that would be even better to do.
 So how common is the problem? And I am going quickly through these because you can look them up on the website, the URL has been provided as quickly as I do. And most of these things are not necessarily something that needs to be discussed in that. How common is the problem? First of all, you would like to have local and current random sample surveys from your area done very recently. Second you would like a second best of systematic reviews of surveys that at least provide information on populations that are similar to the population you are dealing with.
 And third best and local, nonrandom sample. Fourth best case series. Case series consist of 100% of people with the problem. What are you going to say about how common is the problem. And, there is no level 5 for this particular hierarchy.
 No. 2 is this diagnostic or monitoring test accurate? The highest, is in systematic review of cross sectional studies with consistently applied reference standard and blinding. And you will notice that I have highlighted a few words in red. Those are the more subjective terms that are used in these standards that Oxford doesn’t necessarily gives an answer for. So if you ask well, what does consistently apply mean, there is no answer. You have to come up with your own criterion.
 Level 2 is individual cross sectional studies with consistently applied reference standards and blinding. No. 3, nonconsecutive studies or studies without consistently applied reference standards. Nonconsecutive studies Means patients being study without a consecutive series. No. 4 case control studies or poor or nonindependent reference standard and No. 5 essentially there was no studies but reasoning or practitioner's knowledge of how a mechanism works. Prognosis level 1 systematic review of inception cohort studies. Level 2, individual inception cohort studies. Level 3, cohort study or control arm of a randomized trial. Level 4 case series or case control studies or poor quality progress cohort study. And there is no level 5 here.
 Treatment does this intervention help? It should be an extension more or less of Sackett. Level 1, systematic review of randomized trials or N‑of‑1 trials. And there is a plural s here which would suggest that it could be a systematic review of one trial or maybe it is a mistake and level 1 is either you have or had a systematic review of randomized trials or you have an N-of 1 trial that you did on this particular patient. And I presume will have to contact them to clarify that. Level 2 randomized trial or observational study with dramatic effect. No. 3 nonrandomized controlled cohort/follow‑up study. No. 4 case series, case control studies or historically controlled studies and level 5 reasoning based on one's knowledge of the physiology, et cetera.

Common Harms. Systematic very view of randomized trials systematic review of nested case control studies and N‑of‑1 trial, and here it is a single trial, with the patient you are raising the question about or observational study with a dramatic effect. Level 2 individual randomized trial or by exception, exceptional circumstances an observation study with dramatic effect. And I realize now that this ‑‑ observation all study are with a dramatic it could be in level 1 or level 2. That's another one that needs to be clarified.
 Level 3 nonrandomized controlled cohort follow‑up study. For instance, was done in post marketing surveillance, provided there are sufficient numbers to rule out a common harm. Level 4, case series case control or historically controlled study and Number 5. mechanism based reasoning. And lastly, or second to last, Rare harms how to get an idea how common or rare harms you need to study large numbers of cases either in systematic review or you have some other studies that at least included large numbers of subjects.
 So even N of 1 trial one might wonder whether a random sample end of one trial will be sufficient. A level 2 randomized trial or observational study with dramatic effect. Level 3 nonrandomized controlled, et cetera, et cetera. The duration must be sufficiently long. Level 4 case series, case control or historically controlled and level 5 mechanism based reasoning.
 And lastly for screening test ideally randomized review of ‑‑ sorry, systematic review of randomized trials. Second randomized trial. 3, nonrandomized controlled cohort follow‑up. 4, case series, case control, historically controlled studies and level 5 mechanism based reasoning. Now if you paid attention you will have noticed here that there is randomized trials and here is randomized trials. Aren't we talked about a screening test? We will get to that later either today or next week that the optimal screen ‑‑ optimal evaluation of a diagnostic test is to have a randomized trial in which in one arm patients are diagnosed with one test and the other arm they are diagnosed with the alternative test and now we look at what the clinicians do to treat these people and what long term affect is there in terms of the patient's health status. So the best to evaluate diagnostic test is not which one, has the best relationship with a reference test but in the end doesn't make any difference.
 So we will get there in due time. Okay. Here is what I did. My apologies and we will put this one back in this corner. All right called your attention to the subjective elements that were in some of the Oxford language consistently applied, poor standard, poor quality cohort studydramatic effect, sufficient numbers. These are all subjective calls for which you get no guidance. And which make it difficult to apply the Oxford standards, whether that's for systematic reviewers or for people who want to do bedside evidence‑based practice. And we will be talking next week about GRADE and how they don't necessarily solve the problem, but for a systematic reviewer or ask a systematic reviewer A, to make a decision; B, to justify that decision; and C describe how they applied the decision of what, for instance, is in dramatic effect and what's this nondramatic effect.
 Now if you go to the Oxford website, do not get scared that how you see the evidence hierarchies is like this. They are not vertical with one higher on the other. But the six questions are stacked on one another and you have from left to right the range of qualities of evidence. And one of the reasons I think they did this is to emphasize that asking the right question, is just as important as considering the evidence in a hierarchy because the hierarchy as we saw is not necessarily the same for each question. As a matter of fact there are some dramatic differences.
 Okay. There are footnotes to the Oxford 2011 scheme where they say level may be graded down on the basis of study quality which I have been referred to as small d and precision which refers to a wide confidence interval. Indirectness, the PICO of the study does not match the PICO of the clinical question. The studies in a systematic review have either a different population than you want an answer for or a different intervention or a different comparator or a different outcomes than you think you want to have to make a decision on. But it is close enough that you may want to consider if there is no direct evidence. We have more to say about that next week where we talk about GRADE.
 Also Oxford allows grading down because of an inconsistency between the findings of studies, or between the absolute effect size is very small. And also that comes from GRADE and we will pay for attention to next week.
 According to Oxford level may be graded up, if there is a large or very large effect size. If, you treat patients for depression and before treatment they score yet average of 26 on the BDI and after treatment they on average score 29 ‑‑ sorry, 21 points, you have done something good for the average patients but don't flatter yourself. This is just a minor change.
 But if you had managed to reduce average depression levels from 26 to 6 which if I recall well is the average for a nondepressed group, you are a miracle worker! How did you do that?
 You at least looking at the average turned everyone from almost certainly or certainly depressed in to certainly, almost certainly nondepressed. If you can do something that's major, Oxford says and GRADE says you can, grade up the rating of this evidence.
 Lastly Oxford does not offer a screen tooling the quality and quantity of the evidence to the strength of the recommendation. They just ‑‑ let me quickly run back. There is 1, 2, 3, 4, 5 levels of evidence for six questions. Deal with it. How should you act when you have level 1 evidence and how should that be different from what you have level 3 evidence. Oxford does not tell you.
 And I am not sure where that comes from. Now if you go back to the 2009 grid, they have four levels grade recommendation. A is when you get consistent level 1 studies. B is when you get consistent level 2 or 3 studies or you are doing extrapolations from the level 1 studies. C you get when you find level 4 studies or extrapolations from level 2 or 3 studies. And the lowest level recommendation, D is when you have nothing better than level 5 evidence or troublingly inconsistent or inconclusive studies of any level. However they bring you a little bit closer to recommendations but I still didn't find what action they said you should be taking.
 Next we go over to AAN where they resolve this problem for us. In the meantime there should be a slide here reminding me to ask you whether you have any questions about the Oxford Centre for Evidence‑Based Medicine procedures and specifically, you know, what questions they distinguish. And how they rate evidence for those questions. And at least in 2009 linked evidence quantity and quality to grades of recommendation without being specific.
 Now let's look at the AAN manual. That's the American Academy of Neurology and they have what I think is a very beautiful piece of work. A clinical practice guidelines manual and here is the link to their website. The arrow doesn't want to move. But you can clearly see in the yellow box the link. And you can download it for free and it is a pdf. About 70 pages. So let's have a look.
 This is the second edition. The earliest edition was from 2004. And they have done tweaking, some small stuff. They also have introduced some aspects of GRADE in to their methodology. Now supposedly there is to be a 2014 manual which has been finalized by their committee that publishes the manual but the information I have is that the board of AAN hasn't yet blessed the 2014 edition. So we need to deal with the 2011 edition which indeed has to be used by all AAN groups that develop practical ‑‑ sorry, practice guideline manuals or guidelines.
 It is all used by a lot of groups in ACRM, the American Congress of Rehabilitation Medicine who are developing guidelines. Probably its most famous group Cicerone et al, with their three very often quoted systematic reviews doesn't use it. So let's take a quick look at the entire process, before we look specifically at, how does AAN say we ought to handle evidence. They start off with that you first develop the questions that guide your review and your guidelines. They recommend the PICO format or the PICOTS where the T stands for the time frame, how long after onset of treatment do you want to look at outcomes and maybe health care setting, second, what works in socialized medicine in England may not necessarily work in our health care system because it is so different.
 All other questions, whether they concern diagnosis or prognosis, et cetera, are squished in to a PICO, PICOTS framework. For instance, for the screening question, you know, the population is cleared. People you want to screen. The I is doing screening. The C the comparator is not doing screening. The O is the outcome of the screening, what you find, . T is at what point in time, after the screening. Are you looking at outcomes and S, of course, is the second. So there are some discussion of types of clinical questions and then they talk about the development of an analytic framework. And here is an example of a hypothetical and very simple analytic framework that when you develop a guideline, you generally want to talk about everything that has to do with the management and treatment of a particular group of patients. Like everyone who, develops carpal tunnel syndrome or would maybe all the problems or a particular problem that is found in a larger group. For instance, carpal tunnel syndrome in people with spinal cord injury. So in order to do that you need to talk about diagnosis, treatment, prognosis, and you can start talking about an analytic framework where if we have a good test it still might only pick up 90% of the people who have the problem that we are interested in. And then if you have a good treatment, the numbers needed to treat is 4.8 indicating that you need to treat 4.8 patients for every success. And, of course, not a single treatment is good forever. People relapse, et cetera, et cetera.
 So if you have to do a prognosis as to what the situation in five years, you may predict only 30% of your sample correctly.
 So this linking between steps or separate questions and separate systematic review is the analytic framework. And I suggest it may be worth your while to have a look at that.
 Then the next step, of course, is finding and analyzing the evidence. Finding relevant evidence. Two, identifying methodological characteristics of the studies and I turned this red to indicate we are not getting where we want to be.
 Three, rating the risk of bias. Four understanding measures of association. Five, understanding measures of statistical precision. Six, interpreting a study and then C the stage of synthesizes evidence and formulating evidence‑based conclusions. Accounting for conflicting evidence and they had a very nice section about how to know to do that. Knowing how to perform a meta-analysis. Nicely written and not technical at all. Target audience for that is clearly clinicians. Wording conclusions for nontherapeutic questions. Generally therapeutic questions are fairly straightforward but they have some specific guidance when you have questions about diagnosis, prognosis, screening, et cetera.
 And capturing issues of generalizability in the conclusion. Again nicely written. If you have information in the studies that were found, about in slightly different population or slightly different intervention or a slightly different outcome measure, how far can you generalize? Or if you are writing a systematic review what can you suggest how wide this evidence might be used.
 So let's look at some of ‑‑ more on making practice recommendation, rating the overall, confidence in the evidence from the perspective of supporting practice recommendations. Putting the evidence in to clinical context and crafting recommendations. Basing recommendations on surrogate outcomes and knowing when not to make a recommendation and making suggestions for feature research. Some of this we will discuss next week in the framework of GRADE. Others I again refer you to the AAN manual who do a very nice job of this.
 So I pause here to give people an opportunity to type in any questions.

 And it looks like nobody took the pen or the keyboard. So let's move. If you are still formulating your questions, don't hesitate to throw it out later. Sara Woodworth has started typing. Let's see what she comes up which gives me a chance to rest my voice a little bit longer and type a sip of water. Well, nothing is coming forward from Sara. There it is – “what discipline is manual intended, therapist nursing, et cetera?? Well, this is written by neurologists for neurologists and it indeed has a strong medical perspective but there is a lot of simply good thinking in there. If you can put aside the problem of physicians always playing first violin, I would suggest that you, you know, skim or even better read this manual just in the good comprehensive background how responsible attempt at reviewing the literature, finding evidence, using the evidence to phrase recommendations and clinical practice guidelines is done.

I am moving on to AAN details. They have classification schemes for five types of systematic reviewing activities or clinical questions, therapy causation or they refer to well, what do we know what causes this problem.
 This is independent of diagnosing it. It is independent of treating it. But what ‑‑ when before what is the causal network that may exist that explains why some people have this problem and others don't.
 Prognosis, diagnosis and population screening. So it is not quite the same as Oxford's but there is quite some overlap. Let's start with what's easiest to understand, therapy. And AAN and all its five hierarchies distinguishes, defines four levels. High is the best. Sorry. One, Roman I is the best. Roman IV is the poorest. So if you hear me talk about level 2 study, you know by definition that I am talking about the study that gives pretty good evidence but not the top level of evidence. Now for treatment the class I can only be produced by an RCT in a representative population. And again I have highlighted in red terms that not necessarily are defined by the manual. But the manual leaves it up to any particular group that applies its methods to fill in some of these things.
 Relevant baseline characteristics, well, what's relevant? Are presented and substantially equivalent between treatment groups. Or there is appropriate statistical adjustment for differences. This is your famous table I. You have the two columns. One for the treatment group and one for the control group. And we ‑‑ you know that we always hope that those will be pretty much the same between the two.
 Now sometimes you have an unlucky randomization. Even if you randomize large numbers of people you still may have a lucky hand. And, you know, what you can do is in the statistical analysis, use appropriate statistical techniques to balance the two groups out.
 Also acquired in these class I RCTs is concealed allocation, a clear definition of what the primary outcomes are. Clearly defined inclusion/exclusion criteria for the study. And adequate accounting for dropouts with loss to follow‑up no higher than 80%. Sorry, no higher than 20%. . if there are crossovers people who were assigned to treatment but by accident or because they refuse to show up, end up in the other group, the numbers of them are sufficiently small to have minimal potential. The last one is applicable only to noninferiority or equivalence trials. I will not pester you for that. If you are in to those trials you can look it up. We have a big issue with AAN that they specify that I can grab ‑‑ but bullet No. 2 you either have to have masked outcome assessment or objective outcomes. So what is an objective outcome according to AAN? It is an outcome measure that's unlikely to be affected by an observer, the patient, the treating physician, the investigator in a study, expectation or bias. For instance, we could talk about that in the case of blood test or administrative outcome data. Mr. Jones was reported to have died or not to have died. . That or not that is a pretty objective outcome that we don't quibble too much about who is dead and who isn't.
 So they distinguish three levels and maybe the Romans here are inappropriate because it may confuse you with the four levels. But the best is the investigator and the patient cannot influence the measurement of the outcome. . Level 2 is one of them can but not the other. And level 3 is both can impact the measurement of the outcome. And there is specification for AAN guidelines. Only level 1 as defined here is objective. Which of course creates a major problem for disability or rehabilitation studies where we often came out blind. If you want to know whether a particular treatment makes a patient please depressed, pretty much the only way we have to determine whether somebody is depressed or not is to ask them.
 So the patient, is the reporter. Now if you hear that, that's my phone. And somebody is trying to call me. And they will go away in another two rings. If you do behavioral treatment like CBT to treat this patient's depression, there is no way you can hide from the patient that you are treating them with CBT. If you are comparing CBT to some other treatment and you have to get the patient's informed consent to be randomized, probably IRB will ask you to make pretty clear to the patient what the two options are or else he or she cannot make an informed decision whether he wants to be in the research or not. And when you give him or her information on what's involved in CBT and what's involved in the comparison treatment, pretty much as five minutes in to the treatments he or she has randomized to the patient knows that he is in CBT or he knows he is in the comparison treatment.
 No blinding, no blinding. No level 1 outcome. No level 1 outcome, is on this slide. That means by definition bullet 2 here not masked outcome assistance. You cannot be class 1 according to AAN and we will have much more to say about that. But let' goes to class 2. Cohort study meeting the criteria A through E for class I, or RCT that lacks one or more of the criteria B through E in class I.
 In addition to being allowed to miss some of the criteria the person still or the study still shoot report all relevant baseline characteristics and should be substantial equivalence between the treatment groups or appropriate statistical adjustment just like if you have an unlucky hand in class I. And you still should have master objective outcome assessment which means that most of rehab research and disability research cannot even be class II. It ends up in class III. Control studies, including well-defined natural history controls or patients serving as their own controls. Means pre/post studies. A description of major confounding differences between treatment groups that could affect outcome should be presented. And No. 3, outcome assessment is masked or if it is not masked it is objective or if it is not objective, it is performed by somebody who is not a member of the treatment team.
 So now all of a sudden we start talking about an assessor who is not controlled by the treatment team could assess outcomes to still rescue rehabilitation study for class III. Which means my former example of treating depression with CBT versus another never would be level III because the only person who can report on the depression level is the patient and the patient isn't blinded.
 Now there are other studies that we do, for instance, comparing two different ways of teaching a patient how to do self‑care. We can come up with situations where, of course, the therapist know what they are doing. What's the intervention. The patient knows it. But we could arrange for an independent assessor who is blinded, to determine objectively on say ‑‑ defend what the level of the ability of the patient is and in that way one our studies could still be rescued. If that can't be done, our CBT study with patient reported outcome is dumped in to class IV where it is at the same level as garbage studies. Studies that didn't even include patients with a disease who are interested in. Studies that didn't even include the interventions or the comparator that we are interested in.
 Studies that didn't define the interventions or the outcome measures. Or they thought they had great results but you cannot calculate an effect size. Our best finally done, research, by AAN is dumped in to the basement. And we have a lot more to say about that.
 Before I jump to the other hierarchies a quick word about how AAN links the level of the evidence to formulating conclusions, evidence synthesis. Of course, it is ‑‑ it tells you to link the evidence to the clinical question, to consider four types of information. What's a class of information between of all the ‑‑ sorry the class of evidence of all studies that you included in your systematic review. Are they level I, level II, level III, level IV. What effect size did you find between treatment and outcome. And here they took on some aspects of GRADE and may be if there is difference between effect size maybe you want to look only at the effect size of the better study.
 What's the statistical precision of the individual studies and if you go and pool studies, do some meta-analysis, what's the precision of the pooled studies and again this is an aspect of grade they introduced and then what's the consistency between studies. Are they all pretty much pointing to the same outcome when treatment X acts better than treatment Y but some had a pretty good affect size. Some had a minimum effect size but some had a weak affect size. You had some that had no difference and some that had a different and some that found God forbid the comparator was better than the treatment. So all of that ought to be considered when you go to drawing conclusions.
 How does AAN characterize individual studies? One, the highest level give you a lower risk of bias to a moderate risk of bias. Three, a moderately high risk of bias and four, a very high risk of bias.
 Pretty much unless you are interested in doing drug studies, what rehabilitation researchers, disability researchers study in terms of intervention, and with a little bit of work level II or level II, moderately high risk or high risk of bias. Congratulations all of us. Here is the link of the studies to recommendations.
 If you have multiple class I studies, then you should phrase your conclusions something like treatment X is highly likely to be effective. If you have multiple II studies, just a single class I study they say you should phrase it treatment X is likely to be effective.
 If you have multiple class III studies or only a single class II study, treatment X is possibly effective and multiple class IV studies or a single class III study, insufficient evidence in favor or against treatment X. Again going back to my whipping boy, if rehabilitation and disability studies can never be higher than level III, with a little bit of luck and otherwise level II, we will always not get any higher than a recommendation of our treatment is possibly effective.
 There you go. Go home.
 For ineffective therapy the phrasing is more or less systematically with the same differentiation but, of course, reversed treatment X highly likely not to be effective. Treatment X is likely not to be effective and treatment X is possibly not effective and there is insufficient evidence in favor of or against treatment X and for all the conclusions before you can make a statement about not effectiveness, you must make sure that either the individual study has enough power or if you do a meta-analysis, the combined studies have sufficient power or else you are not standing on very solid ground, making a recommendation that uses the word not.
 Okay. Before we go on to some of the other types of questions, any questions that you have for me as to what I just went over with respect to AAN and how it goes about, you know, looking at evidence for treatment questions? If not as usual do not be shy. Type while I go on to the next or type send an e‑mail to Joann after the session and we will try to get you an answer.

 Okay. Let's start looking at prognosis. And I will run through these things fairly quickly because you can all look up the details in the AAN manual and they have examples there. So it gives a good idea of what's going on.
 Class I, you should have a cohort study with prospective data collection. Retrospective, of course, puts you in to the basement. It should involve a broad spectrum of persons at risk for developing the outcome. Your outcome measurement the thing that you are predicting, prognosticating is either something objective or if it is something subjective it is determined, without the person not having knowledge of your risk factor. Your baseline risk factor.
 So it is pretty much knowledge ‑‑ eliminates the patient, because if for subjective outcomes. If what you are trying to predict is whether or not he will be depressed in a year, and if one of your prognostic factors is if he is depressed now, generally patients know whether he is depressed or not of and we will now in a year from now if they are depressed. If a year from now if you ask them if they are depressed they will link that to what they are now. Hmmm. I was depressed a year ago. I must still be depressed now. Or if they don't do that, that's a different story. But that's the without knowledge of risk factor status.
 Also the study, to be qualified as class I should have its inclusion criteria defined and you should have successful follow‑up of at least 80% of the people that you started off at baseline.
 So what about class II? Here can be retrospective data collection. It can be case control study. But otherwise it still needs ‑‑ oh, A and B here. It also needs the broad spectrum of persons and the risk factor and outcome should either be objective things or determined by ‑‑ but without the person who is determining having knowledge of the status at baseline or at baseline having knowledge of the status a year from now, which in the case of retrospective studies is possible.
 Class III cohort or case control study. Narrow spectrum of persons with or without a disease dumped automatically in class III. But you still, need to have an independence of baseline, predictors of risk factors without comes and then, of course, in class IV about anything goes, you may not have people at risk. You may not have people with the risk factor: Undefined unacceptable measures or you cannot calculate an effect size. What you going to do?

Again if you have questions about this, please bring them up. If not, I am going to go on to diagnosis. And I know that for people who, are object rehabilitation disability clinical work or studies you don't do much diagnosing. You probably do a lot more, assessing. So the diagnosis with its yes/no, sick or not sick, it. It probably is not necessarily right up your alley and unfortunately the AAN people do not have an assessment category. But there is still some useful stuff in here.
 So class I, a cohort survey with prospective data collection. Broad spectrum, disease determination is objective or made without knowledge of diagnostic test result. So the people who do the gold standard test, do not know the diagnostic test result and a diagnostic test result do not know the gold standard test.
 Inclusion criteria should be defined and at least 80% of subjects that were started off in the study, which should have the diagnostic test results. And apologies against for my phone ringing. Class II, retrospective data collection is okay but you must meet criteria A and B. You still have to have a broad spectrum of persons and you still have to have making of the diagnosis of the reference test and making of the diagnosis with the new test going on independently from one another.
 Level III cohort case control study or other type study but with a narrow spectrum, you still need independent determination and class IV, everything that is not even good enough for class III.
 Again if you have questions about diagnosis and how it is handled, don't be shy. I will go on to evidence synthesis and formulating conclusions. Which if you are in front of this means that if you noted that I skipped population screening, you are right. And my reasoning was that rehabilitation is people who do disability studies not necessarily in to population screening. But if you are you know where to find the AAN approach.

We talked about this slide and it is true not only as I originally specified it for, evidence relevant to questions of treatment but also to the other questions, prognosis, diagnosis, et cetera. Link it to your critical question. Consider four pieces of information. The class of evidence of all your studies. The strength is the association between the treatment and the outcome. I should have here said between the two parts, of the equation. In some instances this is between the association, between the findings of the diagnostic group and the diagnostic test and the reference test or the strength of the association between predictive factors now and status a year down the road, et cetera. Statistical precision how narrow is the confidence interval around that strength of association. And how consistent are the studies.
 Now if we have conflicting evidence, one study says one thing. Another says another. What should we do? Again AAN gives you some common sense guidance. No. 1, try to explain these inconsistency from the perspective of either systematic error or from random error. You may look at bias, and if there are multiple studies and a weak or broadest among them, C if you get consistency, if you throw out the weak ones, and just look at how the stronger studies do it. And if those are consistent, it is certainly fair to base your conclusion on the studies. Although you out to warn your reader that you threw out some weaker studies because they were contradictory. You can look at random error. . If, you don't necessarily find, that all the weaker studies have one direction or most of them have all the strong studies go the other direction, you can still look at well, if your study is small, it always may come up with the finding that is contrary to the trend. So let's look at what's the power. Are the studies with low power contradictory to the ones with high power? And can I explain stuff that way?
 Or if you do a meta-analysis. If there is sufficient homogeneity of studies. Can you come up with a pooled study effect size with a reasonable confidence interval that will support your conclusion?
 Another approach is to see whether there are systematic or random error or do not so much to strength of the evidence or power but how the studies differ from one another in terms of population, intervention, comparator and outcomes. Are there studies that point one way using these the zoom as a measure of depression and the studies that point another way, use the BDI as a measure of the depression. Well, that might explain it.
 Or the studies that have, you weak effect sizes use the strong comparator while the studies that have strong effect size, use the comparator that really is not anything, that anybody believes should be used with studying with patients because it is such a weak treatment. So there another possibility to explain, differences between study findings based on the nature of the studies.

And a second item that AAN manual talks about is issues of generalizability. Well, you found an effect over a few studies that were done, in the same population with the same intervention, with the same or very close comparators looking at very simple outcomes. Now how far can you apply that result? If we were to apply this to a different population, would that still work if we have studies on stroke. How far can we throw those conclusions when our real population of interest is people with say traumatic brain injury.
 So AAN tells you if you have pretty good findings this far to look very hard at generalizability. And look at what are the subgroups and that may be diagnostic subgroups within that group studied. Or other things, predominately males. Well, and I just looked at the study like that last week, where there were studies predominately males. How far this conclusion works for females? What were the strengths of the doses or the varieties of the interventions studied? What were the comparators? Were the time points after the intervention termination more or less the follow‑ups? pretty much the time and what were the outcomes and limit your recommendation so as to reflect, not reflect it, reflect the limits of the evidence. For instance, it might be it is highly recommended that when women with problem acts receive treatment Y but the effectiveness of treatment Y for men has not been established. So there you take generalizability from to account and you put a break from women to men because based on what you know about men and women with this problem and this treatment, you are not confident that it transfers.
 Or you are just generally very cautious. We already looked at the four interpretations of risk of bias from low through very high bias.
 Well, it is almost our time. I have taken from a study that I have been involved with on application or translation of the criteria that AAN has for treatment that was laid out not by me but somebody else in to a very clear and simple set of steps. And you can link this back to the overview of the criteria for therapies, . What's the design of the study? And it is a simple as if it is randomized control trial it can be class I. If it is not randomized the best is class II. If you have well-defined historical controls, or patients serving as their own controls it is the maximum class III. If you don't have a comparison group, a single subject design, et cetera, et cetera, it is at best in class IV. So put the class A rating, in my arrow again doesn't want to move, but if there to the bottom line in red. You fill it in. And you go to criteria B, which has three sections. Was the person doing the outcome assessment blinded? Yes,: No. Not stated.

And in particular study looked at the fatigue outcome and fatigue by definition is not objective. So we only included questions on the blinding of the outcome assessor. The objective is still in here. Okay. Would have to ‑‑ and the independence of the person doing the assessment. And now based on B1, B2 and B3, if B1 is yes, or B2 is yes, you can be no higher than class I. If only B3 is yes, you cannot be higher than class III. And if all three of them are no or not stated you can be at most class IV.
 Same story for criterion C which already spreads over two pages and then in the end you put it all together what was the rating for a class I. Was that level I, II, III or IV. What was it for B? What was it for C? Now the final is the worst maximum that you had for A, B or C. You pick the worst maximum which should be Roman I, Roman II, Roman III or Roman IV and there you have your risk of bias.
 Again any questions?

If not, we are wrapping up this session. If you have questions e‑mail them to Joann. Her e‑mail address is on here- joann.starks@air.org. If she can't answer it she will pass it on to me. Steve ‑‑ ahh just provided the link to the evaluation. And, of course, you are invited to connect to that link and give feedback on today's session. If you have suggestions for what ought to be covered in future sessions, whether that's session 3 or session 4, or as I indicated previously we are more than happy to cook up either a session 5 or do some more sessions probably after the summer holidays. Please let us know. Otherwise thank you all for being there in today's presentation. Thanks to Steve and Thanks to Ann for hosting the session today. And all of you enjoy employing this new gained knowledge. Follow‑up with getting the AAN manual. Next week we will start talking about, you know, some of this stuff that we already have encountered here either in Oxford or AAN but they stole from GRADE and look at GRADE in larger detail. And then the week after that we will get back to one of the central problems of the day, which is do we have to live if in AAN, the best grade disability rehabilitation research can get is a 3 and more typical grade for it is levels Roman IV or can we do something about that.
 So we will discuss some of that stuff, in the session scheduled for four weeks from today. In two weeks from today GRADE and all its glory. Thanks so much.

 >> ANN WILLIAMS: Thank you so much Marcel, and we appreciate all your insight and we appreciate all your participation today to the participants. Like Marcel said Steven had put up a link to the evaluation. So please fill that out. We really enjoy listening to your input and we appreciate the support of the NIDRR to carry out the webinars. The next webinar in this session will be July 2nd. So we will see you all in a few weeks. Thank you.