


BEHAVIOR CHANGE CONSORTIUM

Transbehavioral Outcomes Assessment Workgroup




Pre-Meeting — July 17, 2002
Washington, DC

FINAL REPORT



The Transbehavioral Outcomes Assessment workgroup would like to thank the following organizations for their support of its activities: The National Institutes of Health, The American Heart Association, The Robert Wood Johnson Foundation.

TABLE OF CONTENTS

AGENDA

Estimated Time	Topic	Presenter(s)
10:30-10:45 a.m.	Welcome and Introductions	Marcia Ory, Lisa Klesges
10:45-11:15 a.m.	Background and Aims · Describe rationale for identifying transbehavioral / clinical outcomes · Specify goals for workgroup meeting	Russ Glasgow
11:15-11:30 a.m.	BREAK	
11:30-12:15 p.m.	Presentation of Issues · Review historical approaches to transbehavioral and clinical outcome assessments · Provide pros and cons of different approaches that could be adopted	Claudio Nigg
12:15-1:30 p.m.	WORKING LUNCH · Informal discussion of additional issues, clarifications, questions from consultants, questions from BCC group, etc.	Geof Williams
1:30-2:30 p.m.	Feedback/Recommendations from Outside Consultants · Reactions to BCC activities · How do these activities fit into greater context of behavior change research activities? · Recommendations to BCC · Identification of strategies for moving field forward	
2:30-3:00 p.m.	Action Steps for Workgroup · Create agenda of future activities · Who will lead the initiative from BCC? · What resources will we need? · Where might we obtain funding?	Lisa Klesges, Claudio Nigg
3:00-3:15 p.m.	BREAK	Refreshments provided
3:15-4:00 p.m.	Future Directions / Wrap-up · Formulate discussion with entire BCC membership · Timeline for initiation and completion of activities · Additional topics for discussion?	Marcia Ory

EXECUTIVE SUMMARY

Established in 2000 as part of the Behavior Change Consortium, the Transbehavioral Outcomes Assessment (TBOA) workgroup developed as a complement to the BCC's Conceptual Mediators and Methodology and Data Analysis workgroups.

The Mission of the TBOA workgroup of the Behavior Change Consortium is to further the science of health behavior change and maintenance through cross-site collaboration. Specfic workgroup goals include development and examination of transbehavioral indices or assessment methods (such as a behavior change index) to be used in behavior change research regardless of behavior being addressed.

With a view to launching the TBOA into an active effort, the TBOA workgroup organized a pre-meeting workshop, entitled "Developing a Transbehavioral Outcomes Assessment: Strategies for Comparing Intervention Success Across Behaviors, Populations and Settings." The workshop was held on July 17, 2002, prior to the bi-annual meeting of the BCC membership July 18-19, 2002 in Washington, DC.

The goal of this workshop is to bring BCC members together with external consultants to address the following objectives:

- Identify the importance and rationale for what the BCC is doing to examine intervention mediators and outcomes across populations and behaviors.
- Present a framework for examining these issues.
- Address the pros and cons of various approaches.
- Seek feedback from outside consultants
- Propose at least one concrete assessment tool to use in cross-site comparisons.

In addition to workgroup members and other select BCC members, four outside consultants were invited to participate in the workshop:

1. **Peter Briss**, M.D., Chief, Systematic Reviews Section, Community Guide Branch , Centers for Disease Control and Prevention;
2. **Steven Belle**, Ph.D., M.Sc.Hyg., Associate Professor of Epidemiology and Biostatistics, Graduate School of Public Health, University of Pittsburgh;
3. **Robert M. Kaplan**, Ph.D., Professor and Chair, Department of Family and Preventive Medicine, University of California, San Diego; and
4. **Helena C. Kraemer**, Ph.D., Professor of Biostatistics in Psychiatry, Department of Psychiatry and Behavioral Sciences, Stanford University.

(A complete list of participants can be found on page xx.)

Baed on several preliminary discussions prior to the workshop, the TBOA workgroup presented nine possible approaches, which fell into three categories. Definitions of each approach can be found in the Overview Slide Presentation on page xx.

1. Behavioral Outcomes
   - Percent meeting criteria
   - Stage approach
   - Average effect size
   - Standardized residual change scores
2. Population Impact
   - Expanded impact equation
   - RE-AIM
3. Clinical Interpretations
   - Average effect size with mortaility weighting
   - Clinically preventable burden
   - Number needed to treat.

The bulk of the discussion following the presentation focused on the pros and cons of each approach, as well as the true objective of this endeavor as well as the approporiate audience for scientific results. A brief summary of each approach follows.

- **Percent meeting criteria.**

  This would provide a simple means by which to compare behavioral outcomes, but would require agreement on a subjective criterion measure for all three behaviors.

- **Stage approach.**

  Stage of change would be a more sensitive approach that the above, but would also require agreement on a definitional criteria. Furthermore, not all BCC interventions are stage-based and such an approach may penalize those without this inherent advantage.

- **Average effect size.**

  Effect size is a flexible statistic by which to describe several outcomes, and can be easily translated into clinically meaningful units. In addition to the invalid practice of extending effect sizes beyond a specific population, this approach assumes that effect sizes for all behaviors are independent and equal and this may not be the case.

- **Standardized residual change scores.**

  In order to utilize this approach, it is neccessary to understand how each behavioral outcome relates to its baseline score. There may be too many confounding mediational variables and/or no homoscedasticity.

- **Expanded impact equation.**

  Due due the variability in recruitment strategies, population retention rates, and definition of control groups, this may biased our conclusions.

- **RE-AIM**

  This approach incorporates several real-world considerations, including individual and organizational factors. Due to its breadth, however, it is also impacted by many of the problems affecting the other approaches.

- **Average effect size with mortality weighting.**

  Although this approach is easily disseminable to public health practitioners, reliance on mortality figures from other studies may dilute our conclusions.

- **Clinically preventable burden.**

  This approach that introduces quality adjusted life-years (QALYs) met with strong support. It may also provide cost-effectivess data that could inform the field. Without complete adherence to an treatment (or control) protocol, it may be difficult to calculate.

- **Number needed to treat.**

  This approach is well suited to individual level interventions, but is not readily adaptable to group interventions. Results may not be generalizable beyond the specific time periods used in each study, and may lose impact when extended to public health trials.

In general, no one approach was favored over the others, however Behavioral Outcomes were thought to be of interest to researchers, Population Impact approaches were considered more accessible to public health interventionists, and those approaches based on Clinical Interpretations were thought to be more meaningful to health practitioners.

Several other approaches were recommended by members of the consulting team, including: a) cost-benefit analysis; b) revisiting mediational analyses as an approach; c) comparison of percentile scoring by rank ordering; and d) the use of recursive partitioning methods to limit subjective judgment of optimal criteria.

There were several suggestions for future consideration and direction. These included:

1. charging various workgroups with the task of analyzing one or more of the suggested approaches. The results will be written up individually, and the collective manuscripts will be published as a special issue of (for example) *Annals of Behavioral Medicine*.

2. Sponsor a small conference to highlight the BCC's agenda in this area.

3. Approach this question as a multi-site prevalence study. Obtain consensus on a continuous measure of behavior change from each workgroup, and produce a "Multibehavioral Prevalence Index Across Populations."

4. Further exploring this issue with other behavioral scientists as a symposium at the Society of Behavioral Medicine's annual meeting.

Next steps in this process were more difficult to define. All agreed that the workshop was worthwhile, and the intellectual exercise was challenging. Most agreed that this endeavor was slightly ahead of the field, and that we would need to solicit more feedback from outside sources. Ongoing funding to support this initiative was also not easily resolved, nor was the questions of human resources once the BCC officially disbanded in Spring, 2003.

It was agreed that the workgroup members would carefully review the recommendations and feedback provided by the outside consultants. The workgroup would continue to meet via conference call to discuss and define future direction(s) before advising NIH program staff and other BCC members about plans to proceed.

LIST OF PARTICIPANTS

Ronald Abeles, Ph.D.
Special Assistant to the Director
OBSSR
Office of the Director, NIH
Bethesda, MD

John Allegrante, Ph.D.
Co-Principal Investigator
Teachers College / Columbia University
New York, NY

Terry Bazzarre, Ph.D.
Senior Program Officer
Robert Wood Johnson Foundation
Princeton, NJ

Steven Belle, Ph.D., M.Sc.Hyg
Associate Professor
Department of Epidemiology
University of Pittsburgh
Pittsburgh, PA

Peter Briss, M.D., M.P.H.
Chief, Systematic Reviews Section
Community Guide Branch
Centers for Disease Control and Prevention
Atlanta, GA

Fernando Costa, M.D.
Staff Scientist
American Heart Association
Dallas, TX

Susan Czajkowski, Ph.D.
Research Psychologist
NHLBI
Bethesda, MD

David Dzewaltowski, Ph.D.
Principal Investigator
Kansas State University
Manhattan, KS

Paul Estabrooks, Ph.D.
Co-Investigator and Assistant Professor
Kansas State University
Manhattan, KS

Larry Fine, Ph.D.
Medical Advisor
OBSSR
Office of the Director, NIH
Bethesda, MD

Russell Glasgow, Ph.D.
Senior Scientist
AMC Cancer Research Center
Denver, CO

Patricia J. (Trish) Jordan, Ph.D.
Cross-Site Communications Coordinator
Behavior Change Consortium
Honolulu, HI

Robert Kaplan, Ph.D.
Professor and Chair
Department of Family and Preventive Medicine
UCSD School of Medicine
La Jolla, CA

Abby King, Ph.D.
Principal Investigator
Stanford UniversitySchool of Medicine
Palo Alta, CA

Lisa Klesges, Ph.D.
Co-Investigator and Associate Professor
University of Tennessee, Memphis
Memphis, TN

Helena Kraemer, Ph.D.
Professor
Department of Psychiatry
Stanford University
Stanford, CA

Claudio R. Nigg, Ph.D.
Assistant Professor
University of Hawaii, Manoa
Honolulu, HI

Marcia Ory, Ph.D.
BCC Scientific Coordinator
Professor, Social & Behavioral Health
School of Rural Public Health
Texas A&M University
College Station, TX

Karen Peterson, D.Sc.
Principal Investigator and
Associate Professor, Nutrition
Harvard School of Public Health
Cambridge, MA

Barbara Resnick, Ph.D.
Principal Investigator
University of Maryland
Baltimore, MD

Susan Solomon, Ph.D.
Senior Advisor
OBSSR
Office of the Director, NIH
Bethesda, MD

Deborah Toobert, Ph.D.
Principal Investigator
Orgon Research Institute
Eugene, OR

Geof Williams, M.D., Ph.D.
Principal Investigator
University of Rochester
Rochester, NY

CONSULTANT BIOSKETCHES

PETER BRISS, M.D.

Chief,
Systematic Reviews Section
Community Guide Branch
Centers for Disease Control and Prevention

Dr. Peter A. Briss is Chief of the Systematic Reviews Section in the Community Guide Branch. The Community Guide Branch is responsible (with a non-Federal Task Force and many partners) for developing evidence-based community practice guidelines. He and his staff conduct and communicate the results of systematic reviews of evidence about community interventions. He also serves as a scientific advisor to the Guide to Clinical Preventive Services, and works with the non-randomized studies group and the public health field of the Cochrane Collaboration. He has worked in a broad range of public health topics including vaccine preventable disease, injury prevention, lead poisoning, and cancer screening.

Dr. Briss began his public health career as an EIS officer assigned to the Tennessee Department of Health and Environment and has participated in public health teaching, practice, and research in Uzbekistan, Russia, and Saudi Arabia. He received his medical degree and residency training at the Ohio State University and is board certified in internal medicine and pediatrics. He has received additional post-doctoral training in applied epidemiology and preventive medicine at CDC and in health management and policy at the University of Michigan. He resides in Atlanta with his wife Susan and their children Erin and Laura.

STEVEN BELLE, PH.D., M.SC.HYG.

Associate Professor of
Epidemiology and Biostatistics
Graduate School of Public Health
University of Pittsburgh

Steven Belle, Ph.D., M.Sc.Hyg. is Associate Professor of Epidemiology and Biostatistics in the Graduate School of Public Health at the University of Pittsburgh. He holds a doctorate from the University of Michigan and a Masters of Science in Hygiene from the University of Pittsburgh, both in Biostatistics. He is the Principal Investigator of a multi-center treatment trial designed to identify racial differences in response to optimal therapy for hepatitis C, and is the Biostatistics and Data Management Core director of the Rheumatic Diseases Core Center.

Dr. Belle also oversees data management and analysis for the multi-center REACH clinical trial, designed to test an intervention to reduce caregiver stress among family caregivers of people with Alzheimer's Disease or a related dementia. In the initial funding for REACH, the 6 sites each randomized participants to different interventions as feasibility studies. The results of cross-site analyses of these studies were important in developing a common intervention that is being tested in the newly funded REACH II project.

ROBERT M. KAPLAN, PH.D.
Professor and Chair
Department of Family and Preventive Medicine
University of California, San Diego

Robert M. Kaplan, Ph.D. is Professor and Chair of the Department of Family and Preventive Medicine, at the University of California, San Diego. He is a past President of several organizations, including the American Psychological Association Division of Health Psychology, Section J of the American Association for the Advancement of Science (Pacific), the International Society for Quality of Life Research, and the Society for Behavioral Medicine. He is currently Chair of the Behavioral Science Council of the American Thoracic Society and President of the Academy of Behavioral Medicine Research. Dr. Kaplan is the Editor-in-Chief of the *Annals of Behavioral Medicine*, Associate Editor of the *American Psychologist*, and Consulting Editor of four other academic journals.

Selected additional honors include APA Division of Health Psychology Annual Award for Outstanding Scientific Contribution, University of California 125 Anniversary Award for Most Distinguished Alumnus, University of California, Riverside, American Psychological Association Distinguished Lecturer, and the Distinguished Scientific contribution award from the American Association of Medical School Psychologists. His public service contributions include various NIH, AHRQ and VA grant review groups, and service on the local American Lung Association (ALA) Board of Directors and the regional research committee for the American Heart Association.

He has served as co-chair of the Behavioral Committee for the NIH Women's Health Initiative, and a member of both the NHLBI Behavioral Medicine Task Force and the Institute of Medicine (IOM) National Academy of Sciences Committee on Health and Behavior. In addition, he is the chair of the Cost/Effectiveness Committee for the NHLBI National Emphysema Treatment Trial (NETT). Dr. Kaplan is the author or co-author of more than a dozen books, and more than 350 articles or chapters.

HELENA C. KRAEMER, PH.D.
Professor of Biostatistics in Psychiatry
Department of Psychiatry and Behavioral Sciences
Stanford University

Helena Chmura Kraemer, Ph.D. is Professor of Biostatistics in Psychiatry in the Department of Psychiatry and Behavioral Sciences, Stanford University. She holds a doctorate in Statistics from Stanford University. Her major interest is, in general, the application of statistical methods to the study of behavior in medicine. Specific areas of interest have included issues related to the reliability of measurement and diagnosis, evaluation of power for statistical hypothesis testing, measurement of effect sizes, risk estimation, moderators and mediators of risk or of treatment in RCTs, signal detection methods for evaluation of the potency of risk factors or the effectiveness of treatments in RCTs. She has written numerous articles on statistical methods, and has served as statistical consultant on numerous research studies, both at Stanford and in collaboration with investigators across the country. In particular, she chaired the Research Steering Committee of the Infant Health and Development Program (IHDP), a national multi-site RCT to improve the health of low birth weight premature infants, and was the statistical consultant to the MTA, a recent national multi-site RCT for the treatment of ADHD children, both of which are considered landmark studies in their fields. She has written two books, one on power and the second on the evaluation of medical tests. She has been elected a fellow of the American Statistical Association, and a member of American College of Neuropsychopharmacology. She was the recipient of the Harvard Award in Psychiatric Epidemiology and Biostatistics in 2001.

OVERVIEW SLIDE PRESENTATION

Transbehavioral Outcome Assessment: Potential Approaches

The Transbehavioral Outcomes Assessment Workgroup of the Behavior Change Consortium

July, 2002

Slide 01

Rationale

- Diseases arise from a constellation of risky behaviors
- Very few individuals do not have multiple risk behaviors
- Multiple-behavior interventions have the potential to have a great impact on public health

Slide 02

Problem

- In order to evaluate multiple-behavior interventions, assessments need to be developed to capture the changes from the different behaviors that allows comparisons of effectiveness across behaviors.

Slide 03

Why a TBOA?

- Intervention - simultaneous or sequential? (TBOA would allow comparison of single-behavior intervention to multiple-behavior intervention using the same metric)
- Identification of key behavioral constructs and processes common across behaviors (i.e., what are the global principles of change?)

Slide 04

Why a TBOA? (cont'd)

- Multiple behavior interactions to increase or decrease health risks
- Relative impact of different interventions targeting multiple behavior change. Which is most effective?

Slide 05

Consideration

- More than one index may be used depending on audience.
- What can the BCC realistically contribute?
- Multibehavioral mediator assessments are also extremely important to identify common mechanisms of change
  - Not specifically presented here, but concepts apply

Slide 06

General Approaches

- **Behavioral Outcomes (BO)**
- **Population Impact (PI)**
- **Clinical Interpretations (CI)**

Slide 07

TBOA#1: % Meeting Criteria

- **Use consensus guidelines**
- **Proportions adoption criteria**
- **Pros**
  - Easy; comparable to existing data
- **Cons**
  - Only credit for meeting criteria; criteria for all behaviors possible?; dichotomize continuous behaviors

Slide 08

TBOA#2: Stage Approaches

- **Ordinal scaling**
- **Pros**
  - More sensitive; can be converted to criteria
- **Cons**
  - Categorical; some interventions are not stage-based

Slide 09

TBOA#3: Avg. Effect size

- **ES for individual behaviors divided by the number of behaviors addressed.**
- **Pros**
  - Continuous; different scales OK; includes studies with different goals
- **Cons**
  - Interpreting to media and policy makers

Slide 10

TBOA#4: Standardized Residual Change Scores

- **Use regression to predict residual change after controlling for baseline levels., then add change scores**
- **Pros**
  - More sensitive; conversion to criteria
- **Cons**
  - Categorical; some interventions not stage-based

Slide 11

TBOA#5: Expanded Impact Equation

- **intervention impact = recruitment x retention x mean efficacy x number of behaviors**
- **The mean efficacy - ES X a coefficient of contribution to all cause mortality**
  - averaged for behaviors addressed

Slide 12

TBOA#5: (cont'd)

- **Pros**
  - Addresses components of an intervention that affect impact on the study sample
- **Cons**
  - What is large impact, may become very small

Slide 13

TBOA#6: RE-AIM

- **Incorporates Reach, Efficacy/ Effectiveness, Adoption, Implementation, and Maintenance**
- **Pros**
  - Considers real world issues of dissemination and successful application
  - Individual and organizational factors
- **Cons**
  - No single metric

Slide 14

TBOA#7: Avg. Effect Size With Mortality Weighting

- **Mean efficacy - ES X coefficient of contribution to all cause mortality**
  - averaged for behaviors addressed
  - could also be done with disease specific mortality or for morbidity outcomes

Slide 15

TBOA#7: (cont'd)

- **Pros**
  - Clinical and public health relevance, context of all cause mortality (or other outcome used)
- **Cons**
  - Multiple behavior impact estimation on outcome?

Slide 16

TBOA#8: Clinically Preventable Burden

- **Definition: Proportion of disease and injury prevented if intervention is delivered to 100% of target population**
- **CPB represented as Quality Adjusted Life Years (QALY)=burden of disease x interventions effectiveness**
- **Cost-effectiveness (CE) costs averted / QALY's saved**

Slide 17

TBOA#8: (cont'd)

- **Pros**
  - Considers quality of life; health outcomes
- **Cons**
  - Assumes complete adherence

Slide 18

TBOA#9: Number Needed to Treat (NNT)

- **Definition: NNT to avoid a single additional adverse outcome**
- **NNT = 1/(ARC - ART)**
- **Pros**
  - Similar to CPB

Slide 19

TBOA#9: (cont'd)

- **Cons**
  - May not apply to different time period than that used in studies; become small in public health trials

Slide 20

Issues to Common Approaches

- **Cost-effectiveness can be incorporated**
- **Do not take into account recycling and/ or relapse**

Slide 21

Specifically for the BCC

- **What can we address specifically with the data collected from our projects?**
  - Mortality, morbidity, cost, dissemination of information, multiple behavior data, etc.

Slide 22

Specific to this Endeavor

- Some agreement is better than none to move the field forward and raise further research questions
- Heterogeneity of outcomes, due to various population and intervention differences may influence strength of contribution to cross-site analyses.
- Logistics of timing to obtain endpoints from various trials.

Slide 23

... this Endeavor (cont'd)

- PI's cooperation to share outcome data – need buy-in.
- Keeping TX assignment blinded until study completion.
- Need to work within the limitations of our data and use this as a stepping stone to propose future research programs to granting agencies.

Slide 24

Final Thought

"Don't worry about people stealing an idea. If it's original, you will have to ram it down their throats."

~ Howard Aiken

Slide 25

CONSULTANT FEEDBACK STATEMENTS

PETER BRISS, M.D.
Chief,
Systematic Reviews Section
Community Guide Branch
Centers for Disease Control and Prevention

"Thou shalt not sit with statisticians nor commit a social science."

~ W. H. Auden

Main take home points were:

1. It is worthwhile to explore whether synthesizing results across studies is feasible.
   a. Looking at the results of multiple studies allows you, in principle, to examine questions no single study can answer, e.g, whether interventions are generally effective or whether they vary based on population, setting, or treatment variation.
   b. Looking at the results of multiple studies may also provide more generalizable results than any single study.
   c. Thinking about synthesis will also improve the individual studies.
2. Apples and oranges comparisons are okay if you're trying to draw conclusions about fruit. Carefully combining information about related but not identical interventions and/or outcomes might help to fully represent intervention and outcome constructs, enhance external validity and usefulness, and identify common aspects of effective interventions (see also 3).
3. Some syntheses will be feasible and useful. Some will not.
   a. "Every data set won't answer every question."
4. To the extent possible, it is good to clarify the goals of summary measures and the audiences for which the measures are intended. These issues will influence the choice of an appropriate measure.
   a. "If you don't know where you're going, any road will get you there."
5. We didn't talk about this much at the meeting, but our group sometimes finds logic models useful in helping people think through what outcome data will be sought, which effect measures will be used, e.g., type of measures, timing of measures, subpopulations, etc., and might be useful for the PI's in planning what data they need.
6. I think that, in general, Claudio's list of potential ways of synthesizing was seen as reasonably clear, complete, and acceptable, as was his list of pros and cons. I think it is clear from the discussion that none of the approaches is clearly right or wrong and that you can afford some flexibility in approaches to transforming and synthesizing (see also comment 4 above).
7. In addition to Claudio's list you could consider percentage change (although this measure is heavily dependent on the measurement of the baseline especially if the baseline is low).
8. The audience was (not surprisingly) a bit split about our favorite measures. In terms of synthesizing across behaviors, there was a general preference for continuous (as opposed to categorical) measures. Standardized mean difference (or equivalent) got a fair amount of attention (which could then be translated to a more interpretable measure for users)
9. NNT got some attention as an interpretable measure for providers although it is harder to use when an intervention has multiple important out-

comes or when the intervention is not delivered at the individual level.

10. Many, but not all, of us would also like to see an exploration of whether you can model downstream consequences of the behaviors. (I personally think that this question is secondary.) Multiple options were discussed including QALYs, cost-benefit analysis, a RE-AIM-type framework, and perhaps an outcome table sort of approach. Some of us cautioned against using this kind of measure be the primary product because you didn't directly measure these outcomes in the studies and you may or may not be the people to do this modeling.

11. If you wanted to collect any additional data here at the end of the process you might consider cost or quality-of-life data (I personally prefer the former).

12. I didn't send PDF files of our papers but people who want to look at how we're been thinking about synthesizing behavioral studies might consider looking at a couple of our papers by Hopkins et al. (2001) and Kahn et al. (2002). They (and other) papers are available at the community guide website at: http://thecommunityguide.org/

References

Hopkins, D.P., Briss, P.A., Ricard, C.J., et al. (2001). Reviews of evidence regarding interventions to reduce tobacco use and exposure to environmental tobacco smoke. [Review] [429 refs]. *American Journal of Preventive Medicine*, *20*,16-66.

Kahn, E.B., Ramsey, L.T., Brownson, R.C., et al. (2002). The effectiveness of interventions to increase physical activity. A systematic review. *American Journal of Preventive Medicine*, *22*,73-107.

STEVEN BELLE, PH.D., M.SC.HYG.
Associate Professor of
Epidemiology and Biostatistics
Graduate School of Public Health
University of Pittsburgh

First of all, I would like to thank you for inviting me to this stimulating session. I found it quite interesting — I thought there were some good ideas expressed and I hope it was useful for those who organized the meeting. My summary comments about the meeting will focus on the main issues we discussed, though I may toss in a few idle thoughts as well.

1. As I interpreted the goal of the meeting, it was to identify methods and measures which could be used to meaningfully compare the effectiveness of multiple interventions that target different outcomes. Before tackling this daunting task, perhaps one should step back and first ask whether such an exercise should be undertaken. If the answer is in the affirmative, a second question may be whether it is possible to attain the goal. I would answer yes to the first question and "I don't know" to the second, but I think that the goal merits further pursuit.

2. Having answered thusly, I'll carry on as if the goal of the meeting should be pursued. However, before addressing issues specific to combining information across studies, it is important to note that having each study meet their goals is paramount, i.e., site-specific analyses need to be completed before one completes anything cross-site. This should be a standard for any attempts to synthesize information across studies. Knowledge is gained about how to combine information when the individual studies are examined. Furthermore, testing their interventions is, appropriately, the primary focus of the investigators. Assuming that that is the game plan, one can start thinking about ways to combine information across sites.

3. The complexity, and what I find to be the interesting aspect of this project, is that the BCC studies are so different, not only the interventions, but also the targets of the interventions. The "art" of

combining data is making use of the commonalities. Without any common ground, it would not make sense to try to combine information from the studies, and I would answer "no" to the first question above. Finding the level at which the studies can be considered similar enough to be combined is what is important. One should be able to make useful inferences about "behavioral" change if one can combine information about different "behaviors" (e.g., diet, exercise, cigarette smoking). There is a caveat, however; when the common ground is found, it may be that the question is no longer interesting, so that combining information across studies is not informative. This would happen, for example, if the particular behaviors were so diverse that combining them under a common rubric was no longer interesting. The behaviors which the BCC interventions are targeting are all associated with (presumably) quality and quantity of life, so that the common ground should be of interest to some.

4. If combining the data seems to be a good idea, then one needs to consider the common ground for numerous dimensions, e.g., what the interventions are, what outcomes the interventions target, and who are in the interventions. In other words, what are the characteristics of the interventions, the outcomes, and the participants that can be examined with an eye toward combining data from multiple studies?

5. Regarding what the interventions are, possibly we can take a page from REACH. There, we decomposed interventions according to a common theoretical construct into what entity is being targeted (caregiver, care-recipient, social and physical environment) and what domains are being targeted (knowledge, skills, behavior, and affect). Since the BCC interventions were required to have theoretical underpinnings, it is likely that such decomposition is possible, at least for those interventions grounded in the same, or similar, theory. For example, though entity may not be relevant, the domain targeted may be one dimension to consider.

6. Regarding the outcomes, the overview prepared by Claudio Nigg seems to be a nice way to organize approaches to combining information from diverse outcomes. There may be other approaches than listed there, but those listed provide a reasonable beginning, and the pro/con points are a useful starting point for discussions. Whatever the summary measure selected, it should be easily understood, have clinical relevance, and meaningfully put all of the outcomes on a similar scale. A couple of the approaches in the document meet these criteria. My personal preference is an approach that incorporates clinical views. Should one be able to determine, for example, the age- and sex-specific number of additional years of life expected associated with reducing cigarette smoking from X_1 to X_2 cigarettes a day, or reducing fat intake from Y_1 to Y_2 %, or increasing METS from Z_1 to Z_2 per week, then I think you have a gone a long way to meeting the three criteria. You can tell an individual person that they have gained so many years of life expectancy and I think that may be understandable and is certainly clinically relevant and puts all the outcomes on the same scale enabling you to compare treatments by some summary measure of number of years of life expectancy gained by those in each treatment group. The drawback may be the lack of available data, but I think it's worth pursuing possible ways to obtain the data (e.g., long-term longitudinal studies such as Framingham, actuarial tables). Of course, the group from which the data are obtained should be as comparable as possible to those in the BCC cohorts, but I think a general population is a good start. People who were recruited into the BCC may have a different life expectancy than members of whatever cohort is used to obtain the estimates, but it still seems to be a reasonable starting ground, particularly if population data are available for the life expectancy estimates.

7. Regarding the participants, there is quite a bit of variability across the sites and this may be the area which requires the greatest amount of consideration concerning how to combine data. For example, there needs to be thought about whether

interventions targeting groups can be combined with interventions targeting individuals. Also, my initial reaction is that sites with participants who do not have common characteristics (e.g., one site restricted to elderly participants versus another site with only school children) should not be combined, but this merits further consideration.

In summary, I think the meeting provided a good basis for further discussions. Site-specific analyses should be paramount, but consideration of cross-site analyses should continue since I believe that pursuing cross-site analyses is worthwhile. This means that common ground needs to be found for a number of dimensions, and possible outcome measures, some of which have been identified, merit further consideration.

ROBERT M. KAPLAN, PH.D.
Professor and Chair
Department of Family and Preventive Medicine
University of California, San Diego

Thank you once again for having me participate in the BCC meeting. I learned a lot and enjoyed participating. My reactions are best summarized as follows.

1. The group should consider estimating public health benefits that result form behavioral interventions. You might consider the recommendations by summarized in the Gold et al book (Cost-Effectiveness in Health and Medicine. New York: Oxford Press, 1996.)

2. Modeling exercises could be very valuable. One example of applications of the a model is provided by Lisa Prosser and colleagues (Cost-Effectiveness of Cholesterol-Lowering Therapies According to Selected Patient Characteristics, *Annals of Internal Medicine*, 2000, *132*, 769-779).

3. You might consider applying a general quality of life outcome measure. I gave Russ Glasgow copies of the QWB-SA. The cheapest and easiest measure is the EQ-5D. It's considerably shorter than the QWB. The difficulty is that it is considerably less sensitive to minor variations in wellness. I would be happy to give you my own (opinionated) evaluation if you would like to discuss it further.

Once again, thank you sincerely for inviting me to the meeting.

HELENA C. KRAEMER, PH.D.
Professor of Biostatistics in Psychiatry
Department of Psychiatry and Behavioral Sciences
Stanford University

1. The first priority, as was recommended, is to complete the studies were designed to do: To provide the answer to the primary question of each study. These should be based on analysis by intention to treat, using the best possible approach in each study. I would recommend against trying to set specific procedures at this stage that must be followed across studies, for that might compromise the quality of the individual studies. The outcome measures should be those set forth in the proposal; the methods should be those described in the proposal unless something is apparent in the distribution that was now known at the time of the proposal and would invalidate the conclusions. In short, do what you said you were going to do, do what your studies were designed to do, before launching into new areas.

2. **Moderators and Mediators:** The definitions of these, and even more so, the statistical approaches, are widely varied and very confusing in the psychology literature. I recommend Baron & Kenny's conceptual definitions (Baron & Kenny, 1986), but my operational definitions (Kraemer, Stice, Kazdin, & Kupfer, 2001; Kraemer, Wilson, Fairburn, & Agras, 2002) . The latter paper (due out in September, but of which Abby has a prepub copy) has very specific instructions on how to use linear models to do the analysis. To see an example of the application see (Group, 1999a, 1999b). The MTA group also has an additional moderator paper in press (JCCP) using ROC methods to find the optimal set of moderators. You could contact Steve Hinshaw at UC Berkeley if you'd like to take an early look at it.

3. **Effect Sizes:** Whatever the outcome, there are many very disparate possible effect sizes. For example, for a binary outcome, you could use risk ratios, risk differences, odds ratios, as well as phi, kappa coefficients of various ilks, gamma coefficients, Yule's Index, Youden's index—no need to go on. What I (and others recommend) is NNT (Number needed to treat to have one extra success or one less failure). If one treated N subjects, the number of successes would be N R_1. If one assigned N subjects to the control group, the number of successes would be NR_0, where the R's are the proportion success in the T and C groups. Thus the number of excess successes would be N(R_1-R_0) with T over C. For 1 excess success N(R_1-R_0)=1, i.e. NNT=1/(R_1-R_0). In this case, the probability that a randomly selected subject from T would do better than a randomly selected subject from C (flipping a fair coin to assign the ties) would be AUC=Prob(T>C)=.5(R_1-R_0+1). AUC=.5 means no difference between T and C. AUC approaching 1 means T much better than C, AUC approaching 0 means C much better than T. Thus NNT=1/{2AUC-1}. Thus if AUC=.5, NNT approaches infinity. If AUC approaches 1, NTT approaches 1.

   The problem is that this is limited to a binary outcome. Where there is a valid and reliable ordinal outcome, it is almost inevitable that there is more power in testing and more precision in estimation in using that ordinal outcome. The more sensitive to individual differences is the ordinal outcome (continuum is better than a 10 point scale, a 10 point scale is better than a 9 point scale etc.), the greater the power and the precision. Some implications: 1) Given a choice between an ordinal outcome and a binary one, always choose the ordinal one for purposes of hypothesis testing and estimation; 2) In making clinical or policy decisions, it is necessary to dichotomize. However, this should be done after hypothesis testing and estimation is done; and 3) Dichotomization for decision making should not be arbitrary and subjective. It should be based on some objective criterion.

   In any case, we need an effect size not limited to binary outcome, or different for different scaling levels, and one (like NNT or AUC) that is easily understood by clinicians and policy makers.

   For example, with outcome measures satisfying

the assumptions of the t-test, the most common outcome measure is Cohen's d (the standardized mean difference between the T and C group means). In this case the AUC=PHI(d/$2^{1/2}$) where PHI stands for the cumulative normal distribution function. Thus if d=.2 (a "small" effect size per Cohen), AUC=55.6%; if d=.5 ("moderate"), AUC=63.8%; if d=.8 ("large") AUC=71.4%. Then conversion to NNT is easy: d=.2 corresponds to NNT= 9; d=.5 corresponds to NNT=4; d=.8 corresponds to NNT=2.

In general, with any univariate outcome measure, one can estimate AUC by taking every possible pair of subjects, one from T and one from C, and scoring the pair +1 if T is better than C, 0 if T and C are equal or indistinguishable, -1 if C if better than T. AUC is then the average of these scores over all pairs. It can be seen that even with a multivariate outcome, as long as one can do valid pairwise comparisons, one can estimate AUC. And as soon as you have AUC, you can estimate NNT. AUC and NNT then are perfectly general effect sizes, applicable over a wide range of outcome measures, easily understood.

Final note: AUC is the effect size associated with the Mann-Whitney test, and can be computed using the U statistic from that test. The above computation, however, makes the meaning of AUC clearer to non-statisticians, I think.

I haven't yet written up these results, but I'll be glad to share them when I get this finished.

4. **Measuring Change versus Status:** See, first of all, Cronbach and Furby for the classical argument (Cronbach & Furby, 1970). The problem when you measure change is that the resulting measure is very sensitive to error in the starting point. This is particularly true when the start is measured by retrospective recall. Cronbach and Furby recommend focussing on the endpoint scores, largely because in their experience starting values are very subject to error of measurement and not very highly correlated with the endpoint score. In my experience, that's not always true. There are many situations in which use of a pre-post change score (NOT using retrospective recall) yields greater power and precision (Kraemer & Thiemann, 1987). So while I'm not a rigid as are C&F about using change scores, I'm very rigid about getting good measures of status (at baseline and endpoint) and assessing the difference between them, rather than getting a measure at endpoint of the change.

   There is always a good argument between use of a difference score, a percentage change score or using the baseline as a covariate (residual change scores). What is best in one situation is not always best in another. Thus if you could show that in both the treatment and control groups, the difference scores are uncorrelated with the baseline, I'd opt for the difference score. If you could show that in both the treatment and control groups, the percentage change is uncorrelated with the baseline, I'd opt for the percentage change. If you could show that in both the treatment and control groups, the relationship of outcome to baseline was linear and with the same slope, I'd opt for residualized change scores. In any other situation, I'd probably go with C&F and recommend using the endpoint measures, remembering that the distribution of baseline scores is the same, given randomization. Then I'd ask whether the baseline is a moderator of treatment (treatment x baseline interaction).

5. **Comments on the approaches in the Nigg MS:**

   a. ***Approach 1:*** Many of the "criteria" used in medical research are very arbitrary. For example, currently the criterion defining obesity is BMI=27. However, when you examine morbidity or mortality outcomes, there does not seem to be any reason for BMI=27. I'm not enthusiastic about any approach that relies so heavily on arbitrary and subjective criteria.

   b. ***Approach 2:*** Similarly, the definitions of "stage" are often arbitrarily set. A great article is Feinstein's on stages of Cancer—which he called the "Will Roger's Effect"

(Feinstein, Sosin, & Wells, 1985). So, same problem as above.

c. ***Approach 3:*** Much better. I like the idea of combining effect size, particularly if that effect size is either AUC or NNT, which then solves the interpretability "cons". However, I'm wary of averaging them. If you average, you essentially assume that all are independent and equally important. That is seldom true. So the question here is how best to combine the effect sizes. If you had an external criterion, you'd be able to use that to help make the decision.

d. ***Approach 4:*** This approach requires that you know how outcome relates to baseline, and that may not be linear, and even if linear, may not have equal variances around the line for all values of the baseline. There are here too many limiting mathematical assumptions to be comfortable for me.

e. ***Approach 5:*** My problem here is that "recruitment" is often a function of what one chooses as the control group, how onerous the measurement schedule it, and other such factors that reflect only the arbitrary research decisions of the study researchers. Similarly retention can sometime reflect efficacy of treatment as well as the skill of the research staff.

f. ***Approach 6:*** I like the concept, but this appears to mix together many very different conceptual issues. I'm not sure how this would help with the problems addressed by the other approaches.

g. ***Approach 7:*** This seems a tough one, since waiting until all subject die to get mortality data is a tough and often impossible task. Using mortality figures from other studies may confuse the issue, since whether the other studies sampled the same populations you did in your various studies would be difficult to guarantee.

h. ***Approach 8:*** Seems more reasonable. However, those who were randomly assigned to treatment (or control) who did not comply with treatment (or control) also have a QALY and costs and effectiveness. From a pragmatic point of view why would you need complete adherence?

6. **Site Differences:** Here "sites" are studies, and differ not only in geographical location (and all that goes with that), but apparently in various details of the design: sampling frame, inclusion/exclusion criteria, type of treatment, assessment battery etc. For that reason, you should never "muddle" the results, i.e., throw data from various sites together and treat the data as if they came from one site. There is a considerable literature on Simpson's Paradox that delineates the types of statistical artifacts that might result. (Bickel, Hammel, & O' Connell, 1975; Blyth, 1972; Hand, 1979; Kraemer, 1978; Paik, 1985; Wagner, 1982).

7. The analysis should be handled via Meta Analysis methods. Whatever the research question, it should be asked separately at each site, and only at those sites where the sample, treatment, assessments are appropriate to ask that question. Thus some sites may be excluded from certain questions. Once the effect is estimated at each site, a test is done for homogeneity across sites. If there is no evidence of heterogeneity (say $p>.10$), the effects can be pooled (not muddled). If there is evidence of heterogeneity, sources of such heterogeneity should be sought.

8. As you know, Steve Belle and I disagree completely on aspects of this issue. I'll let Steve argue his case. However, I feel strongly that the research question is primary. That is, one first formulates a research question, then one tries to find data appropriate to answer that research question. To try to formulate research question to fit the available data is, I think, bound to lead to serious errors. Thus, if the research question relates to "fruit salad", one seeks studies of fruit salad. One does not mix studies, one related to apples, one related to oranges. The answers from

mixing such studies may not relate to fruit salad, any more than to apples or oranges. That is why the criteria for selecting and deselecting studies from the published literature for meta analysis have undergone such scrutiny: Meta Analysis: Meta Garbage?

Back to site differences. The IHDP reference I alluded to is (IHDP, 1990). The above reference to the MTA study also included considerations of site differences. There they found major, strong, main effects of site, but no site x treatment interactions.

REFERENCES

Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182.

Bickel, P. J., Hammel, E. A., & O' Connell, J. W. (1975). Sex bias in graduate admissions: data from Berkeley. *Science*, *187*, 398-404.

Blyth, C. R. (1972). On Simpson's Paradox and the Sure-Thing Principle. *Journal Of The American Statistical Association*, *68*, 746.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—or should we? *Psychological Bulletin*, *74*(1), 68-80.

Feinstein, A. R., Sosin, D. M., & Wells, C. K. (1985). The Will Rogers Phenomenon. *The New England Journal of Medicine*, *312*, 1604-1608.

The MTA Group. (1999a). Moderators and mediators of treatment response for children with attention-deficit/hyperactivity disorder. *1999*, *56*, 1088-1096.

The MTA Group. (1999b). Treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1073-1086.

Hand, D. J. (1979). Psychiatric Examples of Simpson's Paradox. *British Journal of Psychiatry*, *135*, 90-96.

IHDP. (1990). Infant Health and Development Program: Enhancing the Outcomes of Low Birth Weight, Premature Infants: A Multisite Randomized Trial. *Journal of the American Medical Association*, *263*, 3035-3042.

Kraemer, H. C. (1978). Individual and Ecological Correlation in a General Context: Investigation of Testosterone and Orgasmic Frequency in the Human Male. *Behavioral Science*, *23*, 67-72.

Kraemer, H. C., Stice, E., Kazdin, A., & Kupfer, D. (2001). How do risk factors work together to produce an outcome? Mediators, Moderators, and Independent, Overlapping and Proxy Risk Factors. *The American Journal of Psychiatry, 158*, 848-856.

Kraemer, H. C., & Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (in press). Mediators and Moderators of Treatment Effects in Randomized Clinical Trials. *Archives of General Psychiatry*.

Paik, M. (1985). A graphic representation of a three-way contingency table: Simpson's. *The American Statistician*, *39*, 53-54.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician*, *36*, 46-48.

BIBLIOGRAPHY

Blair, S. N., Kohl, H. W., Barlow, C. E., Paffenbarger, R. S., Gibbons, L.W., Macera, C.A. (1995). Changes in physical fitness and all cause mortality: A prospective study of healthy and unhealthy men. *Journal of the American Medical Association*, *273*(14), 1093-1098.

Centers for Disease Control and Prevention. (November, 2000). Measuring Healthy Days: Population Assessment of Health-Related Quality of Life. Atlanta, GA: CDC.

Chattelier, G., Zapletal, E., Lemaitre, D., Menard, J., Degoulet, P. (1996). The number needed to treat: A clinically useful nomogram in its proper context. *British Medical Journal*, *312*,426-429.

Coffield, A. B., Maciosek, M. V., McGinnis, J. M., Harris, J. R., Caldwell, M.B., Teutsch, S.M., Atkins, D., Richland, J. H., Haddix, A. (2001). Priorities among recommended clinical preventive services. *American Journal of Preventive Medicine*, *21*(1),1-9.

Erickson, P., Wilson, R., Shannon, I. (7 April, 1995). Years of healthy life. Healthy People 2000 Statistical Notes, pp. 1-14.

Fiore, M. C., Bailey, W. C., Cohen, S. J., et al. (June, 2000). T*reating Tobacco Use and Dependence. Clinical Practice Guideline.* Rocheville, MD: U.S. Department of Health and Human Services. Public Health Service.

Glasgow, R. E., Lichtenstein, E., Marcus, A. C. (2002). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy to effectiveness transition. Manuscript submitted for publication.

Goldman, L., Phillips, K.A., Coxson, P., Goldman, P. A., Williams, L., Hunink, M. G., Weinstein, M. C. (2001). The effect of risk factor reductions between 1981 and 1990 on coronary heart disease incidence, prevalence, mortality and cost. *Journal of the American College of Cardiology*, *38*(4), 1012-1017.

Hopkins, D. P., Briss, P. A., Ricard, C. J., Husten, C. G., Carande-Kulis, V. G., Fielding, J. E., Alao, M. O., McKenna, J. W., Sharp, D. J., Harris, J. R., Woollery, T. A., Harris, K. W. (2001). Reviews of evidence regarding interventions to reduce tobacco use and exposure to environmental tobacco smoke. *American Journal of Preventive Medicine*, *20*(2S), 16-66.

Jacobson, N. S., Roberts, L. J., Berns, S. B., McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*(3), 300-307.

Kaplan, R. M. (2002) Uncertainty, variability, and the healthcare decision process. In H.S. Friedman and R. Cohen Silver (Eds.). *Oxford Handbook of Health Psychology*. Oxford University Press.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*(3), 332-339.

Kendall, P. C., Marr-Garcia, A,, Nath, S. R., Sheldrick, R. C. (1999). Normative Comparisons for the Evaluation of Clinical Significance. *Journal of Consulting and Clinical Psychology*, *67*(3), 285-299.

Maciosek, M. V., Coffield, A. B., McGinnis, J. M., Harris, J. R., Caldwell, M. B., Teutsch, S. M., Atkins, D., Richland, J. H., Addix, A. (2001). Methods for priority setting among clinical preventive services. *American Journal of Preventive Medicine*, *21*(1), 10-19.

McGinnis, J. M., Foege, W. H. (1993). Actual causes of death in the United States. *Journal of the American Medical Association*, *270*(18), 2207-2212.

Mokdad, A. H., Bowman, B. A., Ford, E. S., Vinicor, F., Marks, J. S., Koplan, J. P. (2001). The continuing epidemics of obesity and diabetes in the United States. *Journal of the American Medical Association*, *286*,1195-1200.

Nelson, D. E., Bland, S., Powell-Griner, E., Klein, R., Wells, R., Wells, H. E., Hogelin, G., Marks, J. S. (2002). State trends in health risk factors and receipt of clinical preventive services among US adults during the 1990's. *Journal of the American Medical Association*, *287*,2659-2667.

Ory, M., Jordan, P. J., Bazzarre. T. (in press). The Behavior Change Consortium: Setting the stage for a new century of health behavior change research. *Health Education Research*.

Raw, M., McNeill, A., West, R. (1999). Smoking cessation: Evidence based recommendations for healthcare systems. *British Medical Journal*, *318*, 182-185.

Rigotti, N. A. (2002). Clinical practice. Treatment of tobacco use and dependence. *New England Journal of Medicine*, *346*(7), 506-512.

Sallis, J.F., Hovell, M.F., Hofstetter, C.R., & Barrington, E. (1992). Explanation of vigorous physical activity during two years using social learning variables. *Social Science and Medicine*, *34*, 25-32.

Stampfer, M. J., Hu, F. B., Manson, J. E., Rimm, E. B., Willett, W. C. (2000). Primary prevention of coronary heart disease in women through diet and lifestyle. *New England Journal of Medicine*, *343*, 16-22.

Taylor, D. H., Hasselblad, V., Henley, S. J., Thun, M. J., Sloan, F. A. (2002). Benefits of smoking cessation for longevity. *American Journal of Public Health*, *92*, 990–996.

Woolf, S. H. (1999). The need for perspective in evidence-based medicine. *Journal of the American Medical Association*, *282*, 2358-2365.

The Writing Group for the Activity Counseling Trial Research Group. (2001). Effects of physical activity counseling in primary care: The Activity Counseling Trial: A randomized controlled trial. *Journal of the American Medical Association*, *286*, 677-687.

Prepared by:

Behavior Change Consortium

Transbehavioral Outcomes Assessment Workgroup

For more information, please contact:

Trish Jordan

Cross-Site Communications Coordinator

e-mail. beatles@etal.uri.edu