STUDY APPROACH

Prior to widespread dissemination of the recommended program, we are proposing a pilot study to assess the implementation process and short range outcomes. A pilot study is a cost-efficient, evidence-based approach to inform decisions about CTSA-wide recommendations and resource allocation as well as any modifications to the SRC Consensus Working Group program. Results will address perceived barriers and facilitators to implementation, impact on the quality and feasibility of clinical protocols, and any effect on the efficiency of the review process. We anticipate that these short range outcomes will have an important longer range impact--an increase in the quality and number of clinical studies completed.

Figure 1 presents the program's logic model (anticipated resources, program characteristics, and desired outcomes), mapped onto three key evaluation questions:

1) Was the recommended SRC program delivered?
2) Was it effective?
3) Was it worthwhile?

The pilot study will directly address the first two questions regarding program delivery and its effectiveness in the short term. Although the pilot study timeline will be too short to systematically measure longer range outcomes, participating CTSAs will be encouraged to track these outcomes beyond the proposed pilot study period.

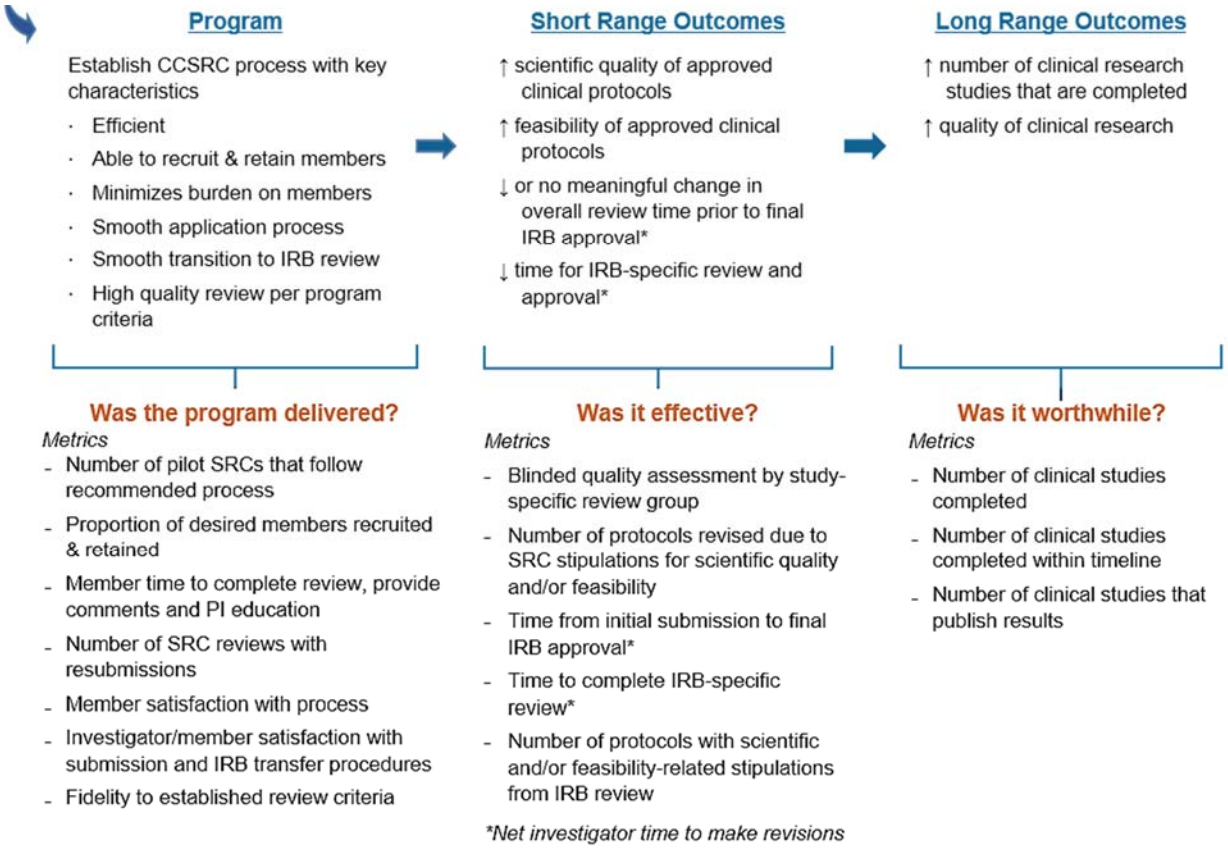
1. MEASURING PROGRAM PERFORMANCE

Quantitative metrics and qualitative interview data will be collected to inform and assess the process of implementation as well as evaluate the program's short-range impact on clinical protocols. This section describes the planned metrics and interview data to address the key evaluation questions. Prior to the start of study data collection, an additional literature search will be conducted to ensure the pilot study incorporates relevant measures from published evaluations of IRB review processes.

a. Process of Program Implementation

As Figure 1 details, an important indicator of the viability of broad dissemination of the SRC Consensus Working Group program is the ability of organizations to implement it successfully with minimal burden. A range of metrics about program delivery will quantitatively assess the implementation process, and a qualitative component will inform perceived barriers to implementation across participating organizations.

Figure 1. SRC Consensus Working Group Program: Logic Model, Evaluation Questions, and Metrics



Quantitative Metrics: The ability to execute the Working Group's recommendations will be measured by the number of pilot study sites that follow the recommendations and the proportion of desired SRC members that sites are able to recruit and retain. Member burden will be measured by the time required to complete a review (including providing comments and education to investigators) and the number of SRC resubmissions. Additionally, members will take part in a short survey to measure their satisfaction with participating, including perceived benefit and burden as well as views on the Protocol Review Form. Additional indicators of program burden will be investigator and member satisfaction with the procedures for protocol submission and the flow of information between the SRC and IRB. Finally, the quality of the SRC review will be measured by fidelity to the review criteria established by the SRC Consensus Working Group. For each protocol, fidelity will be assessed through content analysis of the Protocol Review Form and/or the SRC response provided to the investigator.

Qualitative Component: To understand perceived barriers to establishing the recommended SRC Consensus Working Group program, the pilot study will include a qualitative component prior to implementation. The qualitative component will consist of a semi-structured interview with each site Principal Investigator (PI) about a range of topics related to the recommended

program (Table 1). Results will facilitate more effective implementation during the pilot study and will inform strategies for supporting broader adoption of the recommended SRC processes following the pilot study, if appropriate. In addition, interviews will begin the process of documenting site-level implementation choices about how to implement the program and rationales for those choices.

Table 1. Qualitative Interview Content

Informant	Interview Topics
Site PI	· Anticipated barriers and facilitators · Attitudes about SRC process · Needed supports for implementation · Implementation choices and rationales

b. Short Range Outcomes

Improvements in protocol quality and feasibility without sacrificing efficiency are the predominant markers of a positive impact of the SRC Consensus Working Group program.

Scientific Quality: Consistent with the pre-post design, the primary indicator of changes in protocol quality will be an assessment of a comparable group of protocols sampled during the baseline and implementation phases of the study. Scientific quality will be scored by a study-specific Scientific Quality Review Group charged with rating scientific quality for a random sample of protocols per site. Half of the protocols selected will come from the baseline phase, and half will come from the implementation phase. Scientific Quality Reviewers will be blinded to whether a protocol is from the baseline or implementation phase of the study. To maximize consistency of ratings, the same Scientific Quality Review Group will rate approved protocols across all participating sites using a common rubric comprised of the quality criteria used by SRC members (Objectives, Scientific Merit/Background and Rationale, Design, Eligibility Criteria, Outcomes and Endpoints, Analysis and Sample Size, Data Management). The Scientific Quality Review Group will not include SRC members, but Scientific Quality Reviewers will have the same training as SRC members. The rating system will follow procedures modeled after NIH Study Section processes. That is, to create a numeric score of overall quality for each protocol, reviewers will rate each of the quality criteria separately. Ratings will be averaged to yield an overall protocol rating per reviewer. Each protocol will be rated by two reviewers, whose ratings will be averaged for a final score.

A secondary measure of change in quality will be assessed for each protocol. The number of protocols that are revised in response to SRC stipulations related to scientific quality will indicate a positive change in quality for that protocol. It will be assumed that SRC approval following revisions means that the scientific quality of the protocol has improved; therefore, an independent review of quality for every protocol will not be conducted.

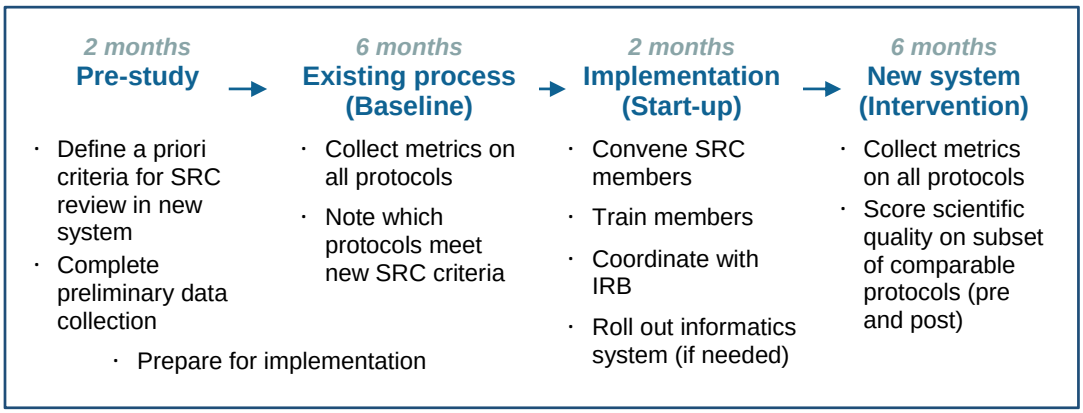
Feasibility: Improvements in protocol feasibility will be measured by the number of protocols that are revised in response to SRC recommendations to

enhance local feasibility. Similar to the per-protocol measure of a change in scientific quality, it will be assumed that SRC approval following revisions means that the local feasibility of the protocol has increased; as a result, an independent review of changes in local feasibility will not be conducted.

Efficiency: To examine whether the SRC Consensus Working Group program meaningfully changes the efficiency of ethical review, the pilot study will collect data on review time and overlap of effort between the SRC and the IRB. Time from initial submission to final IRB approval will address the full ethical review process, including CTSA internal review prior to SRC review (if needed) as well as SRC and IRB reviews. The time for full ethical review will be subdivided into time for CTSA internal review process, time for SRC review, and time for IRB-specific review. Time to complete IRB-specific review will assess whether CTSA internal review and/or SRC review prior to IRB review shortens the IRB review time. All time measurements will be net the turnaround time of investigators for resubmitting revised protocols. Another indicator of efficiency is whether, in practice, SRC review creates overlapping effort with IRB review. The pilot study will address this by measuring the number of scientific or feasibility-related stipulations (the purview of the SRC) that result from the IRB review. The Working Group expects a decrease in these types of stipulations resulting from IRB review after implementing the SRC Consensus Working Group program.

2. STUDY DESIGN

To evaluate program performance, the pilot study will use a pre-post design with baseline data on short range outcomes collected prior to implementation (Figure 2).



a. Pre-Post Comparison

A pre-post comparison has three advantages.

1) All participating sites will implement the recommended SRC Consensus Working Group program and, therefore, will be prepared to continue it after the pilot study, if appropriate.

2) All protocols meeting criteria for SRC review will receive it during the intervention phase. Given the hypothesized benefit of the proposed SRC process for scientific quality and operational feasibility, a study design in which all protocols that meet criteria for SRC review receive it during the study period will accelerate realization of this potential benefit as broadly and quickly as possible.
3) Compared to a design that randomizes protocols, a pre-post design has strong external validity because each site will be implementing only one IRB/SRC process at a time. In contrast, randomization of protocols would require that each site implement both its existing process and the intervention process simultaneously. This would create an organizational burden that would not be present outside of the pilot study.

After considering potential trade-offs of pre-post vs. randomized designs, the SRC Consensus Working Group determined that the advantages of a pre-post design outweigh its weaknesses. First, a pre-post design may introduce a temporal bias if certain types of protocols are more commonly submitted at different times of year. After discussion, the committee concluded that even investigators using the same funding mechanism (e.g., CTSA K awards or pilot study programs) have varying timelines for submitting protocols for review, which lessens the likelihood of temporal bias. Second, although a randomized design could avoid some of the problems presented by confounding variables, non-randomized comparison group designs are frequently used in evaluation studies to assess program impact (Henry 2010). Third, a substantial limitation of a randomized design for this study is that it would not be possible to blind IRB members as to whether a protocol received SRC review because the recommended process calls for IRB members to receive SRC determinations.

b. Multiple Units of Analysis

The study design leverages multiple units of analysis to assess the program. While the intervention will be conducted at the site level, there will be three units of analysis: the site, protocol, and individual. The process of program implementation will be assessed at all three levels. At the site level, metrics will be collected to describe the ability of organizations to implement the program as recommended (e.g., proportion of desired members recruited and retained). Additional metrics will address protocol-level concerns, such as fidelity to recommended review criteria. Member and investigator experience will be analyzed at the individual level, with potential to group by site and baseline categories.

Analysis of short range outcomes will be conducted at the protocol level with pre-post comparison. Pending sufficient similarity across sites, data may be grouped across similar sites to enhance statistical power. The SRC Consensus Working Group considered defining the site as the unit of analysis for short range outcomes, but the number of sites that is feasible for a pilot study is too small for this option.

To ensure comparable groups of protocols at baseline and post-implementation, participating sites will note at baseline which protocols would

meet criteria for receiving SRC review under the new system. This will allow for a two-pronged comparison. The primary analysis group will be protocols that meet criteria for SRC review under the new system during the pre- and post-intervention periods; this comparison will target the effect of the SRC review on directly affected protocols. The secondary analysis group will be all clinical protocols reviewed by the IRB pre- and post-intervention; this comparison will test the more general effect of implementing the recommended SRC process at an organization.

c. Pre-Study Data Collection

Prior to the study, sites will collect preliminary data for two purposes. Determining the volume of activity at participating sites will further inform the number of sites and protocols required for statistical testing. Members of the SRC Consensus Working Group estimated an average of two to twenty-five clinical protocols reviewed by their organizations' IRBs per month, with at least half also receiving SRC review if available. Preliminary data from sites will provide systematic study-specific information upon which to finalize the number of sites required. Additionally, as described above, preliminary data will include 12 interviews with key informants to enhance the team's understanding of anticipated barriers and facilitators, needed supports, site-level procedural details that will usefully inform project planning, and site choices in how to implement the program (see Process of Program Implementation).

d. Incorporating Real-World Variation

The choice of a controlled or usual-practice approach comes with important trade-offs. The results of a study that allows real-world variation could be expected to be more applicable to the full range of organizations, but the measured effect of the intervention could be different than under more controlled conditions. Moreover, with a usual-practice approach, it can be assumed that there will be organizations and conditions under which the intervention will have better or worse outcomes than average, which likely will mute the results. Yet, this also offers an opportunity to discern the interaction of outcomes with different ways the program is implemented.

In order to capitalize on the opportunity to understand how different approaches to implementing the SRC process interact with outcomes, the pilot study will carefully document the key implementation choices at each participating site. The goal will be to identify groups of organizations based on key implementation choices, analogous to identifying different patient groups in an effectiveness trial. As a result, the usual-practice approach of the pilot study not only will provide results reflective of real-world effects but also can provide information about the aspects of implementation at study sites that contribute to the effect's augmentation or mitigation.

3. SAMPLE

a. Recruitment, Sample Size, and Selection Criteria

All CTSAs will be eligible to apply for participation as a pilot study site. To date, a number of CTSAs beyond those represented in the SRC Consensus Working Group have expressed interest in taking part, even without formal recruiting efforts. As a result, we do not anticipate difficulty with recruiting study sites.

The target sample size is 12 CTSAs. Although the sample size may be adjusted pending preliminary data collection, the final sample will not vary greatly in order to ensure that the study remains manageable within the planned timeline.

Selection criteria include 1) documented organizational agreement to participate and 2) the ability to collect the desired metrics. Additionally, the SRC Consensus Working Group recognizes that organizations with CTSAs vary in their existing review processes (i.e., their baseline review systems). To ensure a sample that represents a range of starting points, the study sites will be selected to evenly fill three baseline categories:

1) No scientific review process other than IRB review (n=4),
2) Voluntary scientific review at the departmental or unit level (n=4),
3) Formal scientific review in addition to IRB review (n=4).

Recruiting sites by baseline starting points may also allow analysts to group sites' protocols for increased statistical power (see Preliminary Statistical Power Calculation) and to provide recommendations specific to three categories of existing review systems.

Qualitative Component: A total of 12 key informants (one at each site) will be recruited to take part in a semi-structured individual interview. For qualitative research with a fairly specific research question, a sample of size of 15-20 participants per category typically is sufficient to reach saturation, or the point at which no new information is being gained (Green and Thorogood 2009). While the planned sample size does not reach 15 interviews, interviewing 12 PIs across a range of organizational types produces a reasonable possibility that the qualitative component will reach saturation.

b. Statistical Power Calculation

Based on preliminary data, we expect five protocols per month on average at each site will be eligible for SRC review, and we assume six months of data collection in the pre-intervention period and six months post-implementation. With 12 sites, there will be about 360 protocols in each of the pre-and post-implementation periods. We expect the outcome of mean length of time from submission for ethics review until approval to be 60 to 90 days at baseline (standard deviation (SD) between 40 and 80 days). These time estimates are based on preliminary data and include scientific review but exclude time that the protocol is back in the investigator's hands for revisions. We set the type I error to alpha=0.05, assume the within-site correlation (ICC) to be 0.10, and base the calculation on a two-sided, two-sample t-test. Our assumption about the ICC reduces the sample size of 360 to an effective sample size (ESS) of 92 per time period. If the standard deviation (SD) is 40 days, we will have 80% power to detect a difference pre- vs. post-implementation of 17 days, and if the SD is 80 days, the detectable difference

will be 33 days. If the number of sites were reduced to eight, the detectable difference would be 21 days if SD=40, and 41 days if SD=80.

For the outcome of change in quality of protocols, the Scientific Quality Review Group will assess a subset of 120 protocols total, 60 per time period. The subset for quality review will be further divided by the three baseline categories (no scientific review other than IRB, voluntary scientific review, formal scientific review), with 20 protocols in each of the three baseline categories in each time period. Protocols will be sampled evenly across sites within each baseline category. Again, we set the type I error to alpha=0.05, assume the within-site correlation (ICC) to be 0.10, and base the calculation on a two-sided, two-sample t-test. The ESS is 43 per time period. This will provide 80% power to detect a difference pre- vs. post-implementation of 0.6 standard deviation on the assessment rubric when all 12 sites are analyzed together, and a difference of 1.1 standard deviations for analyses by baseline category. If the number of sites is reduced, we will increase the number of protocols per site to maintain a total of 120 reviews.

4) DATA COLLECTION

Under the direction of the pilot study Project Manager, a staff member at each site will collect and record quantitative data (with the exception of satisfaction surveys). To support site staff in data collection, study sites will be given a 1) data collection guide, 2) pre-study/baseline data collection module based on REDCap platform 3) and recommendations on the informatics system necessary to record study metrics data. The system utilized by pilot sites will ideally serve as both 1) a project management system to support the program workflow (e.g., electronic submission upload and management, time stamping key points in the workflow) and 2) a data management system to centralize additional data entry and quality control. Satisfaction surveys will be fielded electronically by Tufts CTSI.

Qualitative interviews will be conducted by interviewers trained in open-ended, non-leading probing. The Qualitative Lead will develop a semi-structured interview guide, train interviewers, and conduct quality control activities. Since study sites will be located across the country, interviews will be conducted by telephone for cost-efficiency. To ensure data accuracy, interviews will be audio recorded and transcribed verbatim.

5) OVERVIEW OF ANALYSIS PLAN

Analysis will follow a mixed-method framework to integrate findings from quantitative metrics and qualitative interviews (Fetters, Curry, and Creswell 2013; Zhang and Creswell 2012). Findings from qualitative interviews conducted during preliminary data collection will inform quantitative metrics collected (i.e., an exploratory sequential design).

a. Quantitative Metrics

Site and protocol characteristics will be summarized using means, standard deviations, medians, and 25th and 75th percentiles for continuous

and count variables and frequencies and percentages for categorical variables. Program and short-range outcome metrics (see Figure 1) will be similarly summarized, separately for the pre-and post-implementation periods. The short-range outcomes will be compared between time periods using linear, logistic, Poisson, or negative binomial models, as appropriate, adjusting for site and accounting for clustering within site. These models will be further adjusted for key confounders including investigator's level of experience; whether the investigator has submitted multiple protocols in the study period and therefore may have adjusted to SRC requirements over time; and timing of review by a Data and Safety Monitoring Board (DSMB), if applicable, since DSMB review may result in a change to the quality of the protocol independent of the SRC process.

b. Qualitative Interviews

For the qualitative interviews, inductive analysis will begin with consensus coding. To establish a codebook, two analysts will begin by conducting "initial coding" (Lofland and Lofland 1995). During this process, each analyst will independently code a batch of transcripts, meet to compare coding schemas, resolve differences through consensus, and create a codebook with initial codes and definitions. To confirm the codebook, both analysts will apply initial codes to another batch of transcripts and again meet to compare and reach consensus, adding and revising codes as needed. This process will continue until the analysts agree on the codebook and are applying it consistently. At that point, one analyst will code subsequent transcripts, meeting regularly with the second analyst to discuss and resolve coding questions. As coding progresses, analysts will engage in "focused coding" (Lofland and Lofland 1995) by elaborating the codes being used most often and exploring connections between codes. Throughout coding, when a new code is added or changed, previously coded transcripts will be reviewed to determine whether the new code applies. *Atlas.ti* software (ATLAS.ti Scientific Software Development GmbH, Berlin, Germany) will be used to facilitate data organization and coding.

Thematic content analysis (Green and Thorogood 2009) will begin concurrently with coding by using memos to explore connections between codes and emerging themes. In order to fully develop themes, analysts will thoroughly examine all quotations from the most prominent codes. To facilitate comparison by baseline categories in the post-implementation stage, frequencies of the most common codes will be compared across groups. Throughout coding and analysis, analysts will implement procedures to follow Green and Thorogood's general principles for enhancing rigor and credibility: transparency of method, maximization of validity (including attention to deviant cases), maximization of reliability (including frequency counts of themes), constant comparison within the data set and within a case, and a reflexive approach to analysis.

WORKS CITED

Fetters MD, Curry LA, and Creswell JW. 2013. Achieving integration in mixed methods designs-principles and practices. Health Services Research 48 (6 Pt 2): 2134-56.

Green J and Thorogood N. 2009. *Qualitative Methods for Health Research*, 2nd ed. Thousand Oaks, CA: Sage Publications.

Henry, GT. 2010. Comparison group designs. In *Handbook of Practical Program Evaluation*, 3rd ed., Wholey JS, Hatry HP, and Newcomer KE, eds. San Francisco, CA: Jossey-Bass Publishers.

Lofland J and Lofland LH. 1995. Analyzing Social Settings: A Guide to Qualitative Observation and Analysis. Belmont, CA: Wadsworth.

Zhang W and Creswell J. 2012. The use of 'mixing' procedure of mixed methods in health services research. Medical Care 51 (8): e51-7.