Clinical Outcome Assessments (COA) Qualification Program
DDT COA #000019: Skin Infection Patient-Reported Outcome Measure (SKINFECT- PRO)
June 6, 2018 Update

DDT # 0019
June 5, 2018

The ABSSSI instrument has been formally renamed as SKINFECT-PRO© 2017 Foundation for the National Institutes of Health.

Two modifications have been made to the inclusion/exclusion criteria in the updated protocol:

- A footnote has been added to the **fever criteria**, in order to allow for "self-reported history of fever." Sites have reported that patients may have had a fever before being admitted, but the current criteria does not allow for this to be considered an eligible sign/symptom.
- **Expansion of the enrollment window from 24 hours to 48 hours after treatment initiation:** This decision has been made after multiple conversations with the clinicians and site coordinators participating in the validation studies as well as the expert opinion leaders on the FNIH Project team. The narrow window of 24 hours has made it extremely difficult to recruit patients due to the time window of their availability, especially during weekend hours (i.e. admitted on Fri/Sat). We continue to reiterate to sites that they should be enrolling patients as soon as possible; furthermore, during data analysis, we will have the ability to stratify to compare patients who were enrolled 24 hrs vs. 48 hours post treatment initiation.

The contract provides for a step-wise approach for the psychometric evaluation of the content validated instruments and our Team supports that the change to enrolment criteria will not affect the objective of psychometric validation since we are validating the measures prior to evaluating drug effectiveness. This dual approach seeks to spur standardized and harmonized use of the instruments by sponsor as the efforts move along the qualification process.

General Comments:

1. We continue to believe that the current instrument is too long and that it should be further revised to only include items that focus on cardinal symptoms and impacts of the disease and are most relevant to the broader ABSSSI patient population. We are concerned that the current length will increase missing data and limit the assessment's ability to detect change.
   Response: We acknowledge your concern but have decided to move forward with the evaluation of the full items. Indeed, we agree that the current length could increase missing data. However, as a minimum sample size of 75 patients is required to assess ability to detect change in our context and 200 patients will be initially recruited, the assessment of the ability to detect change should not be too limited.

   We recommend that you incorporate a preliminary item reduction phase (including multiple iterations of expert panel consultation and psychometric analyses), prior to instrument finalization, to reduce item redundancy and remove items measuring concepts that can be attributed to treatment side effects (e.g., headache, nausea) or that are distal to treatment effects (e.g., worried). This item reduction phase should be well-documented in your psychometric validation protocol and SAP – including plans for creating an item reduction table detailing your rationale for retaining and deleting individual items and an item-tracking matrix. We recommend that you conduct item selection based on findings from all qualitative and quantitative data analyses and not just Rasch model fit statistics. You may also want to consider a modular approach, where a specific domain (such as cardinal symptoms) with domain based scoring is consider for qualification.
   Response: We agree that Rasch is just one part of the approach to remove items. We have modified the SAP accordingly.

2. Your protocol still lacks details regarding your study administration. Please revise to include additional information about the following:
    a. Data collection procedures for inpatients and outpatients: Procedures will differ for these subpopulations, especially in instances where a patient's condition worsens to the point of hospitalization over the course of the study, following the initial diagnosis of ABSSSI. In these cases, a patient may be enrolled in the study in the outpatient setting, but complete the study in the inpatient setting. Details regarding how these administrations will differ need to be added to the protocol.

Response: We have incorporated details into the protocol, clarifying that data collection procedures are the same for both inpatients and outpatients. Sites will be trained to encourage patients to take their devices to the hospital to continue completing their daily diary if they do become hospitalized. However, if patients become too ill, they will be discontinued from the study and considered a loss to follow-up. Only the data collected up to that point from these patients will be included in the analysis.

3. Please clarify which study protocol is to be followed when the ABSSSI PRO data collection is conducted within the pharmaceutical company sponsored clinical trials, i.e., pharmaceutical company sponsored clinical trial protocol or the ABSSSI PRO psychometric evaluation protocol. In addition, please address the following issues:
    a. The process to ensure the consistency between the pharmaceutical company sponsored clinical trial data collection and the ABSSSI PRO psychometric evaluation data collection.
    b. The process to resolve the discrepancies between the pharmaceutical company sponsored clinical trial protocol and ABSSSI PRO psychometric evaluation protocol, in particular the discrepancy between the inclusion/exclusion criteria and the study design.
    c. The decision making process to stop data collection in pharmaceutical company sponsored clinical trials and to start the data collection from the ICON clinical sites.
    d. The process to pool the data sets collected from different trials.

Response: This additional information will be submitted in the event that this study is implemented within a clinical trial.

4. Information regarding your eCOA system and implementation plan are not included in your protocol. We recommend the following:
    a. Submit Screenshots and training materials (site and patient) for your eCOA implementation for Agency review and comment.
    b. Plan to perform usability testing of eCOA devices and implement a back-up plan (e.g., paper, web-based) in case of any malfunctions with the electronic devices, prior to using the devices in your psychometric evaluation study. Please include details regarding this stage of development and submit protocols and materials related to this usability testing for Agency review and comment.

Responses: A user acceptance testing is currently being conducted on the ePRO devices, but usability testing was not done based on ISPOR's Taskforce's conclusion that such testing is not required for migrations with minor changes. Multiple studies have supported that PRO measures administered on paper are quantitatively comparable with measures administered on an electronic device. [1,2,3] Sites will be trained to tell patients to call the 24-hour CRF Health Help Center if they

[1] Muehlhausen, W., Doll, H., Quadri, N., Fordham, B., O'Donohoe, P., Dogar, N., & Wild, D. J. (2015). Equivalence of electronic and paper administration of patient-reported outcome measures: a systematic review and meta-analysis of studies conducted between 2007 and 2013. *Health and Quality of Life Outcomes*, *13*, 167. http://doi.org/10.1186/s12955-015-0362-x

[2] Van de Looij-Jansen, P. M., & de Wilde, E. J. (2008). Comparison of Web-Based versus Paper-and-Pencil Self-Administered Questionnaire: Effects on Health Indicators in Dutch Adolescents. *Health Services Research*, *43*(5 Pt 1), 1708–1721. http://doi.org/10.1111/j.1475-6773.2008.00860.x

have any issues with their devices. The help center will be able to guide the subject that they must pick up a new device at the site. The help center can help the site prepare a new device for the subject. If the site does not have a device available, the help center can ship the site a new device for the subject. If there is an immediate need for an assessment to be completed, the site can use the web-based platform as a temporary solution to complete the assessments on behalf of the subjects.

5. Currently, your protocol indicates that patients will either complete ABSSSI PRO using the eCOA hand held devices or a telephone interview. We recommend that you move forward with only the eCOA mode of administration (with paper backup in case of device malfunction only) as this will be the least complicated and in alignment with development efforts to date. If telephone interviews are also adopted, you will need to provide details on how patients will be selected for the telephone interviews. Likewise, you will need to develop and submit an interviewer administered version of the ABSSSI PRO instrument (including prompts) for review and comment.
   Response: All US sites are planning to collect data using the ePRO device. In the event that telephone interviews are adopted, an interviewer administered version of the instrument and additional details will be submitted for review.

6. Please provide further details regarding your quality assurance procedures, including: 1) requirements and methods for site and study staff qualifications and training; 2) data monitoring and 3) data entry quality assurance (for paper backup entry into the electronic system).
   Response: Additional details on quality assurance procedures have been added to the protocol. All sites will be trained initially with a web training led by the eCOA team. The eCOA team will then provide an electronic copy of the presentation and site guide to the site, so they have this information is always readily accessible. Study staff will complete a Responsibility Log, which delineates which members are responsible for each study tasks, as well as, sign a Training Completion form which certifies completion of the site training. A customer support center for the ePRO device (CFR Health) is available 24/7 for the sites to call in case they need assistance. All data will be accessible for quality reviewing purposes through TrialManager within one day of data collection. The COA team will review the data every week in batches to ensure that the sites have properly trained their participants to use the ePRO device. If the team sees any odd or missing data, the team will reach out to the sites to investigate and retrain the site if necessary. There will be no paper backup entry as all data will be collected electronically.

7. Please provide details regarding plans for translation and cultural adaptation of the ABSSSI PRO. This instrument will need to be culturally adapted and adequately translated for all intended study populations for use in multinational trials. We refer you to the ISPOR principles for the translation and cultural validation process.
   Response: With regards to the translation and linguistic validation of the instruments, the current contract does not include any costs for this stage. As this instrument is currently under DDT qualification review, the list of countries and languages in which this instrument will be used in is not yet known. At the point of identifying the need for specific language versions of these instruments and securing funding for the translation process, ICON will perform the linguistic validation process adhering to ISPOR's Translation and Cultural Adaptation of Patient-Reported Outcomes Measures-Principles of Good Practice as a guideline. ICON eCOA's current senior scientific consultant, Diane Wild, is the lead author of the ISPOR best practice guidelines for linguistic validation and cross-cultural research methods (2005, 2008) and these papers have been an integral part of the foundation of ICON's Language Services Group. The three instruments will undergo a 10-step linguistic validation process of: preparation, dual forward translation, reconciliation, back translation, back translation review, harmonization, cognitive debriefing, review of cognitive debriefing results and finalization, proofreading, and final report. This linguistic validation process is designed to demonstrate content validity of the translated versions when compared with the source instruments.

[3] Norquist, J., Chirovsky, D., Munshi, T., Tolley, C., Panter, C., & Gater, A. (2017). Assessing the Comparability of Paper and Electronic Versions of the EORTC QOL Module for Head and Neck Cancer: A Qualitative Study. *JMIR Cancer*, *3*(1), e7. http://doi.org/10.2196/cancer.7202

We offer the following additional comments and suggestions related to your submission: ABSSSI PRO Instrument.

We are concerned that Item 23 (daily activities like showering, dressing, or eating) and Item 25 (difficulty sleeping) will not be applicable to the inpatient population as performing these daily activities and sleep schedules would likely be influenced by hospital protocol.

Response: We understand that being in a hospital setting may increase the difficulty for patients to perform daily activities or sleep as regularly as they would have had they been home, but we do not agree that these concepts are "not applicable" to them. We will keep the items and examine the response closely as data is collected.

Psychometric Evaluation Protocol

1. We recommend that you add further details and procedures (e.g., detailed data monitoring at regular intervals; program daily reminders and/or implement daily reminder phone calls or texts for outpatient participants) in order to minimize missing data.
   Response: All data will be accessible for data monitoring through TrialManager (all data will be available within one day of collection). The COA team will review the data every week in batches to ensure that the sites have properly trained their participants to use the ePRO device. If the team sees any odd or missing data, the team will reach out to the sites to investigate and retrain the sites as necessary. The ePRO device has built-in reminders that trigger every day to remind participants to fill out their diary.

2. Please specify whether respondents will be allowed to skip answers or whether each response will be a forced choice. We would prefer if respondents are allowed to skip to avoid erroneous answers. We recommend that you add a skip option to each question and program a logic check that will ask respondents to indicate whether they intentionally skipped items. This way, there is a systematic way to account for missing data.
   Response: We have designed the ePRO to require respondents to respond to every item in order to minimize the missing data. Questions and their understandability have been based on patient content validity and cognitive debriefing interviews and represent simple concepts. Response options also allow for patients to indicate they do not have a given symptom so this would minimize "erroneous' answers. Missing data can have a serious impact on the inferences drawn from a study, and an endpoint may not be evaluable in the event of an unacceptable level of missing data.[4] We understand there is a trade-off between collecting complete but potentially inaccurate data and the possibility of missing data points occurring within a data set that may contain, overall, more accurate data; however, because the items reflect concepts that were selected based on qualitative evidence directly from patients, we believe respondents will be able to provide an accurate answer using the response options.

3. P. 9 notes that the days 7 and 10 administrations will be used to assess test-retest reliability among stable patients based on PGI/PGIC and CGI/CGIC. However, only PGI is available for use to determine stable patients between days 7 and 10 which is correctly stated in P. 20. You may also consider using any two consecutive days' ABSSSI PRO scores from participants whose supplemental question 1 (p. 45) response is "About the same" (this question asks: Overall, how are your skin infection symptoms today compared to yesterday?).
   Response: We have corrected the mistake on page 9 and will be using days 7 and 10 to assess test-retest reliability. We have also included an additional test-retest analysis using supplemental question #1 using consecutive days' ABSSSI PRO scores. The specific analysis to be undertaken will

[4] O'Donohoe, P. Lundy, J.J, Gnanasakthy, A., Greene, A. (2015) Considerations for Requiring Subjects to Provide a Response to Electronic Patient-Reported Outcome Instruments. Volume: 49 issue: 6, page(s): 792-796. https://doi.org/10.1177/2168479015609647

depend on the distribution of responses to supplemental question #1. Should sufficient numbers of patients be stable on across a number of days (e.g., 7), then a mixed model will be used to assess test-retest reliability (ICC) across these days.

4. P. 12: "Specifying a target number of patient subgroups such as inpatients and outpatients is not required for psychometric evaluation": our concern here is whether any items function differently for inpatients than for outpatients. For example, below we raise the question of whether items 24 and 25 mean something different for inpatients and outpatients. In order to address the general possibility of differential item functioning, there needs to be adequate numbers of both inpatients and outpatients. Such numbers may not be reached if target numbers are not set for your sample.

   Response: The SAP has been updated to include target numbers for the three different groups to ensure adequate numbers of inpatients and outpatients have been included.

   P. 14 top: Please clarify the *Known groups validity* section to say that a 150 participants vs. 150 participants comparison is powered to detect an effect size of 0.32, but a 100 participants vs. 100 participants comparison is powered to detect an effect size of 0.18.

   Response: The text has been corrected as follows: "However, a sample of 300 will ensure that the study has sufficient power (80%) to detect, at two-sided $p < 0.05$, differences between two equally-sized study groups (i.e., 150 vs 150) equal to a moderate effect size of around 0.17. For three groups of equal size (i.e., 100 vs 100), a sample of 300 will be sufficient to detect an effect size between any two groups of 0.20 at 80% power."

   P. 14 top: 0.8 should be 0.08.
   Response: This has been corrected.

   P. 17: Please clarify whether the same clinician complete the CGI on days 1 and 14.
   Response: Ideally they should, but this may not be possible and may be a potential study limitation.

   P. 20: Please ensure that Section 5.2 is consistent with the Statistical Analysis Plan document.
   Response: This section is now consistent with the SAP.

<u>Statistical Analysis Plan</u>

1. Section 4.2 Handling of Missing Data
   a. Item and assessment level missingness needs to be assessed. Consider using multiple imputation to handle the missing responses, or consider conducting weighted data analyses with inverse probability of missingness weights. Single imputation with the mean of the observed item responses does not adequately account for variability due to missingness and should be avoided. Additionally, depending on the missingness MCAR may not be a valid assumption. If the MCAR assumption does not hold, then factor analyses and other psychometric data analyses may yield biased results
      Response: While we agree with the statements above, item-level responses will not be missing as the measures will be completed electronically and the system will not allow for skipping of items. Thus, no item-level imputation of missing data will be required. This has been clarified in the SAP. With regard to assessment-level missingness, the SAP stated that LOCF imputation would only be used for assessments in which the participants reach symptom resolution (i.e., for those with a PGI of 'no symptoms today'). On reflection, we now suggest no imputation of missing data. This has been clarified in the SAP.

   b. Using LOCF to impute item responses post symptom resolution: If, after symptom resolution prior to Day 14, participants do not complete daily diaries, then the post- resolution responses are

missing, contrary to the SAP. Instead of handling these responses with LOCF imputation, per the previous two comments there are a number of good reasons why the use of multiple imputation might be appropriate for handling missingness in general, and it could certainly be additionally useful for handling post symptom resolution missingness. In general, LOCF has poor statistical properties, and it is unwarranted to assume that, once rated as resolved, symptoms necessarily stay resolved.
Response: We agree that it is unwarranted to assume that symptoms will remain resolved after the initial rating of symptom resolution. For this reason, we have removed imputation of any missing data from the SAP.

c. The Guidance for Industry: Patient-Reported Outcome Measures (p. 30) recommends at least two sensitivity analyses if multiple imputation is used to handle missing data.
Response: As mentioned above, we will not be doing multiple imputations. This has been clarified in the SAP.

d. Implement procedure to limit the extent of missing responses such as programing daily reminder in the eCOA devices or sending out reminder texts or phone calls.
Response: As mentioned above, the ePRO device has built-in reminders that trigger every day to remind participants to fill out their diary.

2. Section 4.3 Distributional Considerations
   a. Q-Q plots can also be used to assess normality.
   Response: We agree and have updated the SAP to include the generation of Q-Q plots to assess normality.

3. Figure 1: The only arrow coming out of "EFA" should go to "Rasch," and the other arrows currently coming out of "EFA" (e.g., going to "Ability to Detect Change" should instead come out of "Rasch." This is because the proposed Rasch analysis could result in dropping items, and the dropped items would not be included in analyses assessing the ability to detect change.
Response: We agree that it would make more sense to have one arrow coming out of "EFA" going to "Rasch" with the arrows currently coming out of "EFA" to Reliability, Construct Validity, Responder Definitions, and Ability to Detect Change, now coming out of "Rasch." The figure has been revised in the SAP.

4. We recommend that you describe the item selection process in detail and that it be based on results from all analyses including both qualitative and quantitative data.
Response: Yes, we agree and have modified this section in the SAP.

5. We recommend moving up the 5.2.3.1 Item-Level Analysis as the first analysis to be conducted. Specifically, frequency of endorsement, floor and ceiling effect should be examined to inform the factor analysis and Rasch analysis, and whether some items are potential candidates for deletion.
Response: Yes, we agree that the first analysis should be the item-level analysis. This has been modified.

6. Section 5.2.2 EFA
   Given your 2015 final conceptual framework decomposition, please provide rationale for conducting EFA rather than confirmatory factor analysis.

   Response: "EFA will be used to assess the factor structure of the new ABSSSI-PRO instrument. Interrelationships among items will be examined in order to identify the clusters of items that are likely to constitute unique factors (subdomains) within the overall instrument. The EFA will allow the team to identify subdomains. EFA will be used rather than confirmatory factor analysis (CFA) to ensure that the sample size is sufficient to undertake this analysis of domain structure given that estimating sample size for CFA is complex (Wolf et al, 2013)."

7. Section 5.2.3 Rasch Analysis
   a. Please specify the exact Rasch models for polytomous items be used.
      Response: A rating scale model will be used.

   b. If a research category has zero or few responses, it is not feasible to estimate the item parameters. In this case, it may be necessary to use a different day of data for the Rasch model analysis.
      Response: Yes, either a different day could be used, or data could be collapsed across categories.

   c. Please specify whether separate Rasch/IRT analysis will be conducted for each subscale if results from your factor analysis (Section 5.2.2) reveal multidimensionality.
      Response: Yes, a separate analysis will be conducted for each subscale if results from the factor analysis reveal multidimensionality. This has been updated in the SAP.

8. Section 5.2.4 Scoring Algorithm
   a. The SAP describes the scoring algorithm for the "CABP PRO." We assume "ABSSSI PRO" was intended.
      Response: Yes, it should have been ABSSSI PRO. This has been corrected.

   b. Please clarify when an overall ABSSI PRO score will be created in the situation where the ABSSSI PRO is not unidimensional.This will depend on the factor structure and whether, for example, the identified domains are related (i.e., the fit is better on an oblique than an orthogonal rotation)

      Response: This will depend on the factor structure and whether, for example, the identified domains are related (i.e., the fit is better on an oblique than an orthogonal rotation), whether or not an overall ABSSSI PRO score will be created in a situation where the ABSSSI PRO is not uni-dimensional, will depend on the identified factor structure (e.g., whether conceptually it makes sense to sum the domain scores) and whether the identified domains are related. If they are not related (i.e., the fit is better on orthogonal than oblique rotation) then a total score will not be created; if they are related (i.e., the fit is better on oblique than orthogonal rotation) then a total score may be created.

   c. Please clarify how you envision the ABSSSI PRO will be used to define efficacy

endpoints for ABSSSI clinical trials. Relatedly, it is unclear what is the role of a summary score of longitudinal diary data.

Response: We will explore the longitudinal diary data obtained to determine the most appropriate way of defining an efficacy endpoint. It would be best to take scores on any one day, such as baseline and follow-up, rather than averaging over a certain time period, although there may be clinical reasons why it would be beneficial to capture scores over a longer period than one day. We will need to define a responder definition for the ABSSSI PRO e.g. potential endpoint could be achieving the RD, based on scores at baseline and follow-up.

d. The SAP presupposes that classical test theory (CTT) should be used for scoring; Rasch/IRT analysis is intended to play a subsidiary role in determining item adequacy. Please provide rationale the CTT approach is chosen instead of the Rasch/IRT approach to generate scores.

Response: We are recommending the use of CTT rather than Rasch/IRT to generate scores in this instance. We believe that the CTT-based approach is quite adequate for an ABSSSI PRO and that using an IRT-based scoring system may add unnecessary complexity. To this effect we also consider that more data would be required to provide sufficient reliability to generate IRT-based scores.

9. Section 5.2.5.1 Internal Consistency (No comments were included)

10. Section 5.2.8 Exploring Responder Definitions

We agree with the idea of using PGI change and PGIC to anchor the establishment of responder definitions. We suggest basing this on the separate empirical CDFs of ABSSSI PRO change scores within each PGI change and PGIC category. The ROC approach is appropriate to examine the sensitive and specificity of a selected responder definition, however, it is not appropriate for determine a responder definition of meaningful change.

Response: While PGI-C may be sensitive to recall bias, we agree that it is advisable to use both the PGI-C as well as the PGI-S in determining response thresholds, and this has been updated in the SAP accordingly. Our preference, however, is to use ROC methods in addition to other standard anchor-based approaches, and in particular alongside CDF and PDF analysis. We find the results from ROC analysis useful in determining clinically meaningful score change as the optimal/best cut-point specifically identifies the score change on the measure that is best associated with meaningful change on the anchor (e.g. PGI-S and PGI-C). The CDF and PDF plots, by anchor group, provide additional information on meaningful change.

Document updated on 03/18/2019

Statistical Analysis Plan

Psychometric validation of a new SKINFECT patient-reported outcome (PRO) measure Version Number: 3

1. Introduction

Over the past decade, the United States has experienced an epidemic of acute bacterial skin and skin structure infections (ABSSSIs) caused by methicillin-resistant Staphylococcus aureus (MRSA). In addition to Staphylococcus aureus (including MRSA), ABSSSIs are also caused by Streptococcus pyogenes and are among the most common infections encountered in clinical practice.[1] Treatment of ABSSSIs is challenging because of a limited number of safe and efficacious antibacterial medications, especially those administered by the oral route, and the ongoing threat of antibacterial resistance.[2] There is a critical need to develop new and more effective antibacterial agents and for designing clinical trials that can reliably measure treatment benefits that are important to patients, clinicians, and other key decision makers.

In collaboration with the Foundation for the National Institutes of Health Biomarkers Consortium (FNIH BC), ICON has created a new patient-reported outcome (PRO) measure, titled the SKINFECT-PRO ©[3], to assess ABSSSI symptoms and impacts in a clinical trial setting. Having evaluated the content validity of the measure in qualitative interviews, the next step is to validate the measure psychometrically in line with FDA guidance (FDA, 2009). The purpose of this statistical analysis plan (SAP) is to provide full details of the statistical analyses that have been outlined in the study protocol. The scope of this plan includes all the proposed analyses to be executed by ICON. The SAP outlines the rationale for the statistical tests that will be performed and the criteria that will be used to interpret the results. A list of tables summarizing the analyses is also provided. Associated statistical programming code for undertaking these analyses will be prepared by ICON based on the SAP.

2. Objectives

The objective of this study is to evaluate the psychometric properties of the new SKINFECT-PRO© measure, which will be administered as a daily diary. The psychometric properties of the SKINFECT-PRO© will be measured in a patient population characterized by cellulitis (including erysipelas), abscesses, and wound infections (traumatic or surgical site). This is part of a broader effort between ICON and FNIH BC to support an FDA label claim submission used in clinical trials for anti-bacterial interventions and other studies as appropriate. The psychometric properties the study will assess include:

- Item level properties Domain Structure
- Reliability Construct validity
- Ability to detect change
- Responder definition

Document updated on 03/18/2019.

[1] Deleo, F.R., Otto, M., Kreiswirth, B.N., & Chambers, H.F. (2010). Community-associated methicillin-resistant Staphylococcus aureus. Lancet, 375:1557-68.

[2] Foundation for the National Institutes of Health Biomarkers Consortium Project Team. (2011). Recommendations to the FDA for Interim Endpoints for Clinical Trials in Acute Bacterial Skin and Skin Structure Infections (ABSSSI Docket ID: FDA-2010-D-0433).

[3]