Clinical Outcome Assessments (COA) Qualification Program
DDT COA #000018: Pneumonia Patient-Reported Outcome Measure (PNEUMO-PRO)
October 19, 2017 Update

FDA Comment

We acknowledge that you will be proceeding with your psychometric study prior to item reduction. However, it is possible that administering the draft instrument items prior to item reduction phase can impact the sensitivity of the instrument and its ability to accurately assess core concepts of CABP. In the absence of a formal item reduction phase prior to psychometric testing, we recommend that you engage in multiple iterations of item reduction using qualitative (i.e., expert consensus panel review with subject matter experts, including FDA representatives) and quantitative methods to ensure that the most relevant items are included in the final instrument. Note that another clinical study may be needed to confirm the psychometric properties of the reduced instrument. Additionally, you will need to provide greater detail on your ePRO implementation plan, instrument administration schedule, and proposed analyses in your next submission.

Response from the ICON/ FNIH team

Thank you for your comments on our submission.

- Multiple iterations due to "absence of a formal reduction phase": All items included in the current instrument were generated from the evidence from qualitative interviews with patients in the content validity stage, in which patients stated that the included concepts are important to them in pneumonia. The psychometric validation study will allow us to perform an item-level analysis that will evaluate based on evidence whether an item should be removed. We do not want to risk reducing the items and eliminate concepts that patients stated are important in pneumonia based on the content validity assessment previously performed.
- Need for another clinical study: We do not anticipate another study will be needed at this stage as we have kept the instrument item's coverage broad. If further data collection suggests that we need to change existing items or add in new ones, we would pursue a follow up clinical study. We are aware that following item reduction, theoretically, items can perform differently in the absence of other items, but in practice we do not believe the impact of this is large enough to warrant conducting a further clinical study.
- ePRO implementation plan: The requested details have been provided below and have been updated in the protocol.

After inputting the additional clarifications, we seek to move forward with submitting the updated protocol to our sites' internal IRB committees for approval in order to prevent delays in study start-up. We would like to propose a meeting between our respective teams of statisticians in order to reach a consensus about the Statistical Analysis Plan as soon as possible.

FDA Comment (General Comment #1)

We recommend that you focus on items 1-7, which constitute core symptoms of CABP. A thorough review by the project team of the Biomarkers Consortium of the FNIH found support for a symptom improvement efficacy endpoint based on these cardinal symptoms that can be used in a noninferiority trial as a part of the primary efficacy outcome. FDA concurred with this approach and found historical evidence for a treatment effect and noninferiority margin that

supports the selection of an early symptom improvement efficacy endpoint for the noninferiority trial. Any deviation from these cardinal symptoms has potential to alter the assay sensitivity and therefore create difficulty for the use of a noninferiority trial to establish efficacy of a new antibacterial drug for treatment of CABP. The remaining sections may be considered for use as part of supportive endpoints.

Response from the ICON/ FNIH team

We anticipate that findings of the psychometric validation will ultimately lead us to focus that includes the core symptoms mentioned above. However, per FDA's Guidance for Industry on PRO Measures (2009), we will wait until evidence from the psychometric validation supports item reduction.

We believe that capturing all relevant concepts should improve, not diminish, assay sensitivity. Members of this study team were part of FNIH team that evaluated the proposed symptoms submitted to the docket for the CABP guidance. The symptoms and response options suggested by the FNIH team at that time were based on review of recent non-inferiority studies and historical literature, not randomized placebo/no specific therapy controlled superiority trials as would normally be done to evaluate effects of a control drug to establish assay sensitivity, and were based on expert clinician consensus without the benefit of input from patients as required in the FDA PRO guidance. The plan as spelled out by this same FNIH team and submitted to the FDA docket was to use the 7 symptoms as a placeholder until patient interviews could be performed. The 7 symptoms have not been "validated" as measuring or establishing assay sensitivity or as a measure of all relevant concepts in pneumonia. The 7 symptoms were not evaluated for content validity based on patient input. The plan specified in the FNIH submission to the docket, supported by FDA members on the FNIH study team, was to proceed with appropriate evidence-based development of a PRO, using the 7 symptoms while the PRO was developed.

The additional concepts captured were listed by patients themselves as important concepts in CABP and indicate that the initial chosen 7 symptoms lack content validity in that they do not comprehensively represent patients' symptoms of pneumonia. An indication of "treatment of CABP" should address *all* relevant symptoms of the disease. It is common that patient interviews capture more symptoms than clinician interviews and research shows that patient captured symptoms often are more important and more reflective of patient benefit.[1,2] The instrument would be used for all types of trials and studies, not only noninferiority trials. Sponsors could choose to measure subsets of symptoms should they desire to do so, but would have to address why they would not measure symptoms that patients stated were important in the disease.

FDA Comment (General Comment #2)

Your protocol still lacks details regarding your study administration. In your next submission you should include additional information about the following:

---

[1] Justice, A. C., Rabeneck, L., Hays, R. D., Wu, A. W., & Bozzette, S. A. (1999). Sensitivity, specificity, reliability, and clinical validity of provider-reported symptoms: A comparison with self-reported symptoms. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, *21*(2), 126-133.

[2] Justice, A.C., Chang, C.H., Rabeneck, L. & Zackin, R. (2001) Clinical importance of provider-reported HIV symptoms compared with patient-report. Center for Research on Health Care, VA Pittsburg Healthcare System, Section of General Internal Medicine, University of Pittsburgh, Pennsylvania 15240, USA *Med Care* 39:397-408.

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 or hospitalization with ventilation over the course of the study, following the initial diagnosis of CABP. 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 when the setting changes will differ need to be added to the protocol.
b. Exit interviews: Exit interview procedures lack detail and it is unclear whether accommodations will be made for non-English speaking Hispanic/Latino patient populations. Details regarding any accommodations need to be outlined in the study procedures.

Response from the ICON/ FNIH team

a. We will incorporate details to clarify 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 or need to be ventilated, 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.
b. The aim is to conduct a total of n=10 exit interviews with English-speaking patients of Hispanic/Latino ethnicity. We do not currently have the instrument available in other languages, and therefore the PV will only focus on English-speaking patients. Because the sample size is small, we do not anticipate difficulty in finding English-speaking patients on Hispanic/Latino origin.

FDA Comment (General Comment #3 a-b)

Information regarding your ePRO system and implementation plan is not included in your protocol. We recommend the following:

a. Submit screenshots and training materials (site and patient) for your ePRO implementation for Agency review and comment.
b. Plan to perform usability testing of ePRO 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.

Response from the ICON/ FNIH team

a. Please see the attached subject guide, screenshot documents, and training materials.
b. A user acceptance testing was 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. [3, 4, 5]

---

[3] 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

[4] 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

[5] 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

Sites will be trained to tell patients to call the 24-hour CRF Health Help Center if they 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.

Question for the FDA: Do we need to schedule a teleconference to discuss the materials and receive approval before our sites move forward with IRB submission?

FDA Comment (General Comment #4)

Currently, your protocol indicates that your mode of administration will either be ePRO or a telephone interview. We recommend that you move forward with only the ePRO 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 CABP PRO instrument (including prompts) for review and comment.

Response from the ICON/ FNIH team

All US sites are planning to collect data using the ePRO device. Data collection through telephone interviews are only mandatory for sites in Mexico, as this was deemed the better method based on prior experience with similar studies. Spotty and unstable data coverage on personal handheld/ePRO devices made the data transfer unpredictable and unreliable, resulting in missing data and frustration for participants, as internet coverage in Mexico is available for approximately 20% of participants based on prior experience. In prior studies there were no differences between phone and electronic administration of questions. The CRF Health website is designed to be used during the phone interviews, allowing for answers to be entered directly onto the web.

Furthermore, we have not yet received the FDA's approval on our strategy to combine HABP/CABP recruitment. Once we receive this approval, we will submit an interview administered-version of the CABP PRO instrument.

FDA Comment (General Comment #5)

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 from the ICON/ FNIH team

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. Please refer to the attached training presentation and site guide.

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 2 weeks 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.

FDA Comment (General Comment #6)
Please provide details regarding plans for translation and cultural adaptation of the CABP 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 from the ICON/ FNIH team
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.

Our founder, former Vice President and 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.

FDA Comment (CABP PRO Instrument #1)
For items 24-29, we recommend removing the "Not Applicable" response option. We don't believe that "Not Applicable" is a meaningful response option for these items (e.g., Item 24 – "Did you have difficulty sleeping?") and it is unclear how these options would be scored. Additionally, we are concerned that Item 24 (difficulty sleeping), Item 25 (difficulty doing your usual activities), and Item 27 (social activities) will not be applicable to the inpatient population as level of independence (doing usual activities, social interaction) and sleep schedules would likely be influenced by hospital protocol.

Response from the ICON/ FNIH team
Can you please clarify this comment? It is a bit unclear whether your team is recommending the removal of the "Not Applicable" option or requesting validation of its need as one of the response options. The addition of the "Not Applicable" option was informed by the qualitative interviews. To clarify, Question 24 ("Did you have difficulty sleeping?") does not actually have a Not Applicable option; however, questions 25-29 (difficulty doing your usual activities or getting around; difficulty doing daily activities like showering, dressing, or eating; ability to participate in social activities; feeling upset; feeling worried) do have one as necessitated by the qualitative evidence generated from the cognitive debriefing interviews with patients. If we find that we do not receive any N/A answers for these questions, we will be able to address this during evaluation of the scoring algorithm and item evaluation.

In terms of scoring, we proposed standardizing and then summing domain items in the SAP (see SAP Section 5.2.4). We have adjusted this section to include calculating score means rather than sum scores in order to account for any N/A responses.

FDA Comment (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 from the ICON/ FNIH team
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 2 weeks 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 attached screenflow images and PRS touch android document will be used to train all sites and participants. The ePRO device has built-in reminders that trigger every day to remind participants to fill out their diary. Participants who will be responding (sites in Mexico) via telephone interviews will receive a call every day from the sites, so no additional reminders are thought to be necessary.

FDA Comment (Psychometric Evaluation Protocol #2)
P. 11 – Please specify whether respondents will be allowed to skip answers or whether each response will be 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 from the ICON/ FNIH team
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.[6] 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. (Note that we have included the "Not Applicable" as an option for some of the items in the optional impact domains.)

FDA Comment (Psychometric Evaluation Protocol #3)
P. 16 – An error was found. Please correct "0.8" to "0.08."

Response from the ICON/ FNIH team

[6] 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

This error has been updated in the protocol.

FDA Comment (Psychometric Evaluation Protocol #4)
P. 17-18 – There is discrepant information regarding your test-retest reliability analyses. You initially indicate that scores from days 7 and 10 will be used to assess test-retest reliability. However, on p. 18, you state that scores from days 7 and 14 will be used. Please clarify your analysis plan and correct the protocol and SAP accordingly. Please also consider using consecutive days' CABP PRO scores from participants whose supplemental question 1 (p. 55) response is "About the same" (this question asks: *Overall, how are your pneumonia symptoms today compared to yesterday*?).

Response from the ICON/ FNIH team

- We have corrected the mistake on page 18 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' CABP PRO scores. The specific analysis to be undertaken will 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.

FDA Comment (SAP #1 a-f)
1. Section 4.2 Handling of Missing Data

a. Three studies are referred to in your description of test-retest reliability ("Test-retest reliability for all three psychometric evaluation studies"). Please specify what three studies you are referring to.

b. 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.

c. If items are not reduced, participant burden will be high given the frequency of administration and you might have increased levels of missing data due to respondent fatigue. In order to increase your power, we ask that you consider the use of multiple imputation beyond handling missingness at baseline.

d. 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, we recommend that you use multiple imputation to handle post symptom resolution missingness. In general, LOCF has poor statistical properties, and it is unwarranted to assume that symptoms will remain resolved after the initial rating of symptom resolution.

e. 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.

f. For all case report forms (CRFs), we recommend that you add a field at the top of each page to where patient personal identification numbers can be inputted by site staff or pre-populated.

Response from the ICON/ FNIH team

a. The studies refer to the CABP, HABP, and ABSSSI studies. However, this statement has been clarified in the SAP as, "Test-retest reliability for this study […]"

b. 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.

c. Refer to the response above regarding item-level missing data. While we appreciate the need to reduce items and limit participant burden, 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.

d. 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.

e. As mentioned above, we will not be doing multiple imputations. This has been clarified in the SAP.

f. All CRFs will be electronic, so there is no need to add a field at the top of each page for patient IDs. All devices will be linked to a patient ID.

FDA Comment (SAP #2)

Section 4.3 Distributional Considerations

We recommend that you also consider generating Q-Q plots to assess normality.

Response from the ICON/ FNIH team

We agree and have updated the SAP to include the generation of Q-Q plots to assess normality.

FDA Comment (SAP #3.a-b)

Proposed analysis order of operations (Figure 1): We recommend that you modify the figure as follows:

a. Have only one arrow coming out of "EFA" going to "Rasch"
b. Remove other arrows currently coming out of "EFA" (e.g., going to "Ability to Detect Change") and have these instead come out of "Rasch." This is because the proposed Rasch analysis could result in item deletion and these deleted items would not be included in your analyses to assess the ability to detect change.

Response from the ICON/ FNIH team

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.

FDA Comment (SAP #4.a-b)

Section 5.2.3 Rasch Analysis

a. Please indicate which of the two Rasch models for polytomous items (partial credit model or rating scale model) you intend to use.

b. Many items on Day 1 may have no or few responses of "Not at all," which would compromise the accuracy of parameter estimates. Consider whether there is value added to use not only Day 1 but also other days' data in your analysis. 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. Also, if your factor analysis reveals multidimensionality, indicate whether a multidimensional Rasch/IRT analysis will be performed

Response from the ICON/ FNIH team

a. We are intending to use a rating scale model. This has been clarified in the SAP.

b. We agree that there would be value in using days' data other than Day 1 in the Rasch/IRT analysis given the likelihood of no or few "Not at all" responses. We will run the analysis on a number of days to explore item performance on different days. This has been clarified the SAP.
While we could do this, we are currently not intending to run a multidimensional Rasch/IRT analysis if more than one factor is identified on factor analysis. Instead, we are intending to run the analysis on separate factors to assess the unidimensionality of these individual factors. This has been clarified in the SAP.

FDA Comment (SAP #5. a-b)

Section 5.2.4 Scoring Algorithm

a. If your factor analysis reveals multidimensionality, you should consider whether it is appropriate to compute an overall CABP PRO score.
    a. Please clarify how you envision a CABP PRO total score will be used to define efficacy endpoints for CABP clinical trials.

b. The SAP presupposes that classical test theory should be used for scoring; Rasch/IRT analysis is intended to play a subsidiary role in determining adequate items. Please clarify why you are not using Rasch/IRT analysis to generate scores.

Response from the ICON/ FNIH team

a. We appreciate the need to justify the calculation of an overall CABP PRO score. The results of the factor analysis may help to clarify whether a total score is appropriate by examining the relative size of the factor Eigenvalues, percent variance explained, and whether or not an orthogonal or oblique rotation is necessary to obtain a satisfactory solution (if an oblique rotation is necessary then this suggests the existence of a higher level factor). A hierarchical factor analysis could also be performed to confirm the presence of an overall factor, and thus the appropriateness of an overall CABP PRO score. This has been clarified in the SAP.
    a. As we do not expect there to be a total score, the domain scores will be used in the analysis to define the efficacy endpoints.

b. We are intending to use Rasch/IRT analysis to identify, along with the results of classical test theory, the appropriate items to include in the measure. We are not assuming that Rasch/IRT analysis has a subsidiary role here; if anything, it has the stronger role. This has been clarified in the SAP. We do not believe, however, that there will be much advantage from using Rasch/IRT analysis to generate scores compared with the standard classical test theory approaches of taking sum/average item scores. There may be advantage, but we believe that larger sample(s) would be required for these Rasch/IRT scores to be reliable enough to roll out in scoring algorithms. However, based on the results of the study, we will decide if we have enough data and whether it is appropriate to use IRT to generate scores.

FDA Comment (SAP #6.a-c)

Section 5.2.5.1 Internal Consistency

a. You propose that Cronbach's alpha be estimated based on the Day 1 sample. We recommend that Cronbach's alpha be estimated at several different time points in order to capture possible changes in alpha over time.
b. It might be useful to recreate Reeve and Fayers' Figure 1.5.4 using your IRT analyses to examine the reliability of the CABP PRO at different levels of health status.
c. If multiple subscales are defined we recommend that you evaluate internal consistencies separately for each subscale, in addition to the total score.

Response from the ICON/ FNIH team

a. We agree, as with our response to question 4.b above, that there would be value in repeating our analyses on a number of different time points in order to evaluate whether Cronbach's alpha changes over time[7]. This has been updated in the SAP.
b. Yes, we agree that it is important to show the reliability, and specifically the information obtained from the CABP PRO at different levels of health status, and this has been updated in the SAP accordingly[8].
c. This is our typical approach to assessing the reliability of subscales, and we have clarified this in the SAP accordingly.

FDA Comment (SAP #7.a.i - ii)

Section 5.2.8 Exploring Responder Definitions

a. Your proposed use of the Patient Global Impression of Severity measure (PGI-S; referred to as PGI in your submission) in determining response thresholds for the CABP PRO instrument is acceptable. However, we recommend that you use both the PGI-S and PGI-C to provide an accumulation of evidence to help interpret a clinically meaningful score change in the CABP PRO instrument using anchor-based methods and cumulative distribution function (CDF) analysis. We prefer this approach over using ROC curves.
    i. We recommend that you consider using the Clinician Global Impression of Severity (CGI-S) and Change (CGI-C) measures as supportive evidence to help further bolster the patient observations. Both the patient-rated and

[7] Biemer, PP, Christ, SL, Wiesen, CA. A General Approach for Estimating Scale Score Reliability for Panel Survey Data. *Psychological Methods* 2009 December; 14(4): 400-412.
[8] Reeve, BB, Fayers, P. Applying Item Response Theory Modelling for Evaluating Questionnaire Item and Scale Properties. In P. Fayers & RD Hays (Eds.) Assessing Quality of Life in Clinical Trials: Methods of Practice (2nd ed.). New York; Oxford University Press: 2005.

clinician-rated anchor scales should be assessed at the same time points as, but completed after, the CABP PRO instrument.

ii. You should generate CDF plots depicting changes in the CABP PRO score(s) by corresponding patient and clinician global impression of change and severity item response options (i.e., separate curves for each response option).

Response from the ICON/ FNIH team

a. 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 alongside CDF 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).

i. We will use the CGI-C and CGI-S as supportive evidence of the results from the PGI-S and PGI-C and this has been included in the SAP.

ii. We agree that the generation of such CDF plots would be important and we have included production of these in the SAP.

I. Introduction

The Foundation for the National Institutes of Health Biomarkers Consortium (FNIH BC) is interested in developing reliable, well-defined and clinically relevant endpoints that measure tangible benefits for patients in clinical trials of antibacterial drugs. FNIH BC identified Community- acquired Bacterial Pneumonia (CABP) as a priority indication, and subsequently developed a candidate list of endpoints for use in clinical trials. As part of this effort, the FNIH BC seeks to develop a patient-reported outcome (PRO) symptom instrument in accordance with the Food and Drug Administration (FDA) PRO guidance used to support labeling claims (FDA PRO guidance, 2009) for use in clinical trials of antibacterial interventions. The intention is that the PRO instrument will be used to identify and assess symptoms related to clinically relevant endpoints for CABP.

The ICON Commercialisation & Outcomes (ICON) Clinical Outcomes Assessments (COA) group is collaborating with the FNIH BC to develop three reliable, well-defined, and clinically relevant PRO symptom instruments, that measure tangible patient benefits of treatment interventions in antibacterial drug clinical trials, one in CABP, one in hospital-acquired bacterial pneumonia (HABP), and one in acute bacterial skin and skin structure infection (ABSSSI). Through a consortium-based approach, the FNIH BC Project Teams, together with ICON COA, have applied symptom related evidence generated from the published literature and results from qualitative, post-treatment interviews to create the current CABP and ABSSSI disease models and conceptual frameworks. This work informed the development of the proposed CABP-specific PRO and ABSSSI-specific PRO instruments for use in future clinical trials of antimicrobial drugs. Using the same approach, work is currently underway to develop a PRO instrument for hospital-acquired bacterial pneumonia.

The FNIH BC has requested the new CABP PRO, ABSSSI PRO, and future HABP PRO instruments be developed according to the FDA qualification process outlined in the FDA Qualification Process for Drug Development Tools Guidance (Qualification Process DDT Guidance, 2014). This protocol details the objectives, methods, and analysis required for ICON to demonstrate the psychometric properties of the CABP PRO in accordance with the FDA PRO guidance, and satisfy the communication and scoping document requirements for the qualification process. A separate protocol has been prepared for ABSSSI and will be developed for HABP in the near future.

II. Project Objectives

The objective of this study is to evaluate the psychometric properties of the new CABP PRO instrument. The psychometric properties of the CABP PRO will be measured in a patient population characterized by a diagnosis of CABP. This project 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 that the study will assess include:

- Item level properties (item variability, item-total correlations, Rasch analyses)
- Domain Structure (Exploratory Factor Analysis (EFA))
- Reliability (internal consistency, test-retest)
- Construct validity (known groups/discriminant, convergent/divergent)
- Ability to detect change
- Responder definition (distribution-based, anchor-based)