individuals who indicate they have no chronic health condition. Results are compared using mean errors (ME), root mean squared errors (RMSE) and the proportion of values estimated within [0.05]. RESULTS: Using an age adjusted baseline, we found the additive (and multiplicative) methods underestimate the majority of HSUVs (ME:0.0784±0.0254; RMSE:0.1012±0.0651; 26±(6%<−0.05) while the minimum (and ADE) overestimate the majority of HSUVs (ME:−0.0995±0.0695); RMSE:0.1214±0.0950; 20±(33%<−0.05). Although the simple linear model produced the most accurate results (ME:0.0001; RMSE:0.0598; 63±<−0.05), there were some substantial errors with 20% of errors greater than the minimum important difference (0.074). When subgrouping by actual HSUV (range 0.30±−0.917) we found the magnitude and direction of errors in the estimated HSUVs are driven by the actual HSUVs being estimated in addition to the technique used. In general the HSUVs of an adjusted baseline were more accurate than the estimated ones using a baseline of perfect health. CONCLUSIONS: This study makes an important contribution to the evidence in this area as it is the first to compare the five different techniques in the same data set. While the simple linear model gave the most accurate results, the model requires validating in external data and additional research exploring an alternative model specification is warranted.

ISSUES IN THE TRANSLATION AND LINGUISTIC VALIDATION OF CAREGIVER RATING SCALES REGARDING THE BEHAVIOR AND DEVELOPMENT OF CHILDREN AND YOUNG PEOPLE

OBJECTIVES: Caregiver rating scales, intended to evaluate the behaviour of children and young people, are frequently used in clinical trials involving youngs. However, the translation and linguistic validation of such scales can be problematic due to the differing cultural markers of behaviour and development. This study aims to document the problems that can occur, with the hope of facilitating future studies and producing guidelines to avoid cultural compromises when such measures are developed.

METHODS: Past Oxford Outcomes projects, which included the translation of Caregiver Reported Outcomes, were evaluated to identify problematic items. These included the Vineland-II and ABAS (behaviour development scales), ELIDQOL (epilepsy and QoL scale) and WEFERS (functional impairment rating scale) among others. RESULTS: Numerous cultural and linguistic issues became apparent, including the following:

- Many examples of sports and activities were used in the documents, which required thorough cultural adaptation, e.g., types of games.—Logistical cultural differences were made apparent, e.g., caregivers mention children’s understanding of specific concepts or traffic signals, which vary culturally.—Some documents involved markers for identifying speech development, such correct use of irregular verbs. These were problematic in other cultures and speech development specialists were required to find suitable alternatives.—
- More idiomatic expressions are used in PROs developed for adults, e.g., “on the go”, these cause difficulties in translation.—Items surrounding activities such as housework or helping look after siblings are not equivalent in some cultures due to differing role expectations. CONCLUSIONS: The validation of caregiver reported outcomes through interviews with caregivers was particularly important with these scales to ensure cultural appropriateness in target languages. Physician or specialist input was sometimes required to find culturally relevant alternatives. When such measures are created, culturally specific markers of behaviour should be avoided if possible.

RECOMMENDATIONS ABOUT TRANSLATIONS IN THE FINAL FDA GUIDANCE ON PRO MEASURES: WHAT HAS CHANGED AND WHAT HAS REMAINED

OBJECTIVES: Almost four years were necessary to develop the final FDA guidance on the use of PRO measures in clinical trials. Our objective is to compare how the recommendations about translation and cultural adaptation evolved from the 2006 draft to the 2009 final guidance. METHODS: Both guidelines were retrieved through a search on the FDA website and analyzed. RESULTS: Structure and content were modified. Recommendations on translation and cultural adaptation were moved to another section within the EPRO Instruments Part: from “IVD. Modification of an existing assessment tool” to “IVD. PRO Instruments for specific population” for specific content. Furthermore, the text in the body of the final guidance is more concise compared to the draft. The novelty lies in the stipulation that the FDA will review the process used to translate culturally adapt the instruments. As a consequence, an appendix (section 8.8) was added in which the FDA explains which topics should be addressed in the documents provided to the FDA for review: description of process used, patient testing, rationale for decisions, copies of versions and evidence about validity. They are however key points which did not change: the need for providing evidence that content validity as well as other measurement properties are similar between all versions are similar between all versions and the expectation of developing the final guidance on the content to be provided to the FDA. The last point of the Appendix is debatable as we anticipate that it might add a burden in term of costs to provide evidence about the psychometrics of all versions.

SYSTEMATIC REVIEW OF THE RESPONSIVENESS OF SF-36 HEALTH SURVEY MEASURES TO EFFICACIOUS PHARMACEUTICAL THERAPIES IN WELL-CONTROLLED CLINICAL TRIALS

OBJECTIVES: To determine how often SF-36 Health Survey measures respond to efficacious pharmaceutical treatment benefits in well-controlled clinical trials. METHODS: We conducted a systematic review of randomized, double-blind, placebo-controlled trials published in 124 journals in 1995 through 2009 documenting differences between treatment groups for primary medical endpoints and any of the SF-36 component summaries, or eight subscale scores. Concordance was defined in terms of agreement: an adjusted baseline SF-36 endpoint was more accurate than the estimated HSUV. CONCLUSIONS: This study makes an important contribution to the evidence in this area as it is the first to compare the five different techniques in the same data set. While the simple linear model gave the most accurate results, the model requires validating in external data and additional research exploring an alternative model specification is warranted.

DOES DATA COLLECTION FROM ONLINE COMMUNITIES RESULT IN BIASED RESPONSE?

OBJECTIVES: Although the ability to interact with patients in an on-line environment has expanded substantially over the past few years, many researchers are concerned that participants may not be representative from a medication experience perspective (i.e., biased towards complainers). The purpose of this study is to investigate patient responses on treatment satisfaction using a validated PRO measure, the Treatment Satisfaction Questionnaire for Medications (TSQM), collected through a survey of patients with depression from an on-line community. METHODS: A random sample of 500 iGuard.org members treated with an antidepressant were invited to complete an online version of the TSQM, a widely used validated 14-item generic treatment satisfaction instrument. iGuard.org is an online patient community that provides a free medication monitoring service to patients. Non-parametric item response analyses of raw scores revealed that individual items of the TSQM discriminated differences in patient satisfaction. That is, total scores increased the probability of low scores on the individual items decreased and the probability of higher scores increased. As expected, patient satisfaction was related to reported side-effects, with those reporting side-effects experiencing lower satisfaction with medication than those without reported side-effects. CONCLUSIONS: The results from this analysis suggest that PRO survey data collected from a random sample of members of the on-line patient community iGuard.org can be representative of the spectrum of anticipated treatment satisfaction responses. Continuing to explore the potential of direct data capture from on-line patients will be important as researchers seek faster and cheaper alternatives to traditional physician-based recruitment.

A COGNITIVE DEBRIEFING METHODOLOGY FOR ESTABLISHING EQUIVALENCEDURING E-PRO MIGRATION

OBJECTIVES: Electronic data capture, without signifi cantly altering format or text, qualifi es as a minor modifi cation not requiring a full validation (Coons et al. 2009). However, this does not mean that the two formats are perceived in the same way by patients. We aim here to describe a methodology successfully used to establish equivalence between paper and electronic PROs. METHODS: To demonstrate the equality of these different forms of EPRO, we have used a combination of “think-aloud” and retrospective cognitive debriefing techniques, as well as usability testing. The debriefing exercise is designed to assess whether the electronic device changes the way respondents interpret the questions or response options. The usability testing assesses ease of use and identifies issues that may prohibit the use of the ePRO by the target