HEALTH ECONOMICS STUDY GROUP

TITLE: DEVELOPMENT OF THE EXACT-U: A PREFERENCE-BASED ALGORITHM TO REPORT COPD EXACERBATION UTILITIES FROM THE EXACT

Authors: Jennifer Petrillo, John Cairns

Health Services Research Unit, London School of Hygiene and Tropical Medicine

Background: Evidence suggests the EQ5D may be unable to detect differences in all COPD and exacerbation severity levels. There is a need to develop a measure that captures the full exacerbation with precision and responsiveness. The EXACT is a condition-specific tool to measure COPD exacerbations.
Objective: To develop and validate the EXACT-U, an algorithm to report utilities from the EXACT, for use in cost-effectiveness studies in the UK.
Methods: EXACT items were reduced using Rasch analyses. Items and levels were grouped to form health states and cognitively debriefed with patients to ensure appropriate wording. Members of the general public valued 11 randomised health states each (including best/worst states) using TTOs from full health/dead over 10 years. Analyses include OLS regressions to develop models to best predict utilities. Models assessed against actual utilities in the validation sample by number of inconsistencies predicted, R^2, MSE, and RMSE. Preferred model validated using a separate dataset to test responsiveness and discriminant validity.
Results: Five items with 3 to 5 levels were selected. 60 health states were developed for valuation. TTO interviews conducted with 400 respondents: 36 yrs old (13.5 SD); 39.2% male, and 46.4% White British. Preferred model had no predicted inconsistencies; R2=0.334; MAE=0.239; RMSE=0.346. Responsiveness by standard response mean was 0.430 (Day 4), 0.620 (Day 8), and 0.696 (Day 13). Discriminant validity demonstrated among mild, moderate, severe, and very severe exacerbations and stable COPD.
Conclusions: The EXACT-U is a condition-specific preference-based measure to measure patient utility change in COPD exacerbations from baseline through to resolution.

Key words: EXACT-U, health state valuation, condition-specific, time-trade-off, COPD, chronic bronchitis, emphysema

Corresponding author:
Jennifer Petrillo
PhD Candidate
Health Services Research Unit, Department of Public Health and Policy
London School of Hygiene and Tropical Medicine
Keppel Street, London WC1A
Jennifer.Petrillo@lshtm.ac.uk

I. INTRODUCTION

Reimbursement agencies conduct health technology assessments based on cost-effectiveness in terms of impact on health related quality of life (HRQL), and reported in the form of QALYs, or quality-adjusted life years. Interventions that increase QALYs, either by improving the state of health by utilities or increasing life expectancy, are more likely to be reimbursed than those demonstrating no difference. If there is no difference in utility, payers may conclude there is little value in approving or offering an intervention from an economic perspective, particularly if it is more expensive than current treatments.

Respondents value health states by determining to what extent they prefer living in the described state.[1] Utilities are reported either by directly valuing health state descriptions or by asking respondents to complete a questionnaire with a pre-determined value set for the health states. The latter method refers to preference-based measures, which are generic or condition-specific instruments used to report utility scores of a patient when marking their own health.

Generic preference-based measures, also called multi-attribute utility measures (MAUs) for their multi-dimensional composition, evaluate HRQL across a number of dimensions that are designed to apply to a wide range of conditions. In order to standardize utility measurement and reporting for health technology assessments, the EQ5D was selected as the preferred utility instrument for UK cost-effectiveness studies.[2] The extent to which the health states reflect the full experience of a condition, such as COPD, depends on the accuracy of description and range of dimensions addressed. Thus, conditions that affect dimensions not captured in the MAU may not be adequately measured in order to reflect accurate health state change.

As use of the EQ5D has increased, data has emerged suggesting it is unable to detect change associated with effective interventions in certain disease areas, including asthma, urinary incontinence, visual acuity loss, heart failure, and sexual functioning.[3-8] One such study showed that a COPD-specific instrument was able to reach statistically significant differences between moderate and severe COPD where the EQ5D was not.[9] This lack of discrimination in COPD severity from the EQ5D has been reported in two other studies.[10, 11] Evidence also

suggests the EQ5D may be unable to detect differences in COPD and exacerbation severity levels or to demonstrate adequate responsiveness to interventions, presumably due to the limited dimensions on the instrument.[9, 12] Additionally, for exacerbating patients, EQ5D values may not always be collected at the appropriate time to capture the patient's full exacerbation experience due to administration timing. Poor administration timing could lead to incorrectly measured health status change, which could conclude a lack of overall treatment benefit in many cases.

Utilities reported from a condition-specific measure have the potential to maintain sensitivity to patient change and to differences in severity levels by focusing on domains that are relevant to the condition at hand. A number of studies recently have used condition-specific measures to report utilities after developing a preference-based scoring algorithm.[5, 13-15]

The Exacerbation of Chronic Pulmonary Disease Tool (EXACT) is a daily diary designed to measure the frequency, severity, and duration of an exacerbation.[16] A preference-based algorithm could report utilities of an exacerbation from the EXACT. Utility values produced may be used to show the HRQL impact at onset, resolution, or for the duration of an exacerbation. These values could be used in lieu of EQ-5D values as a more detailed and therefore more sensitive measure of COPD exacerbations and potentially of COPD. They could also be used to complement the current EQ-5D values of stable COPD patients when reporting for the condition as a whole.

The objective of the study was to develop and test a preference-based algorithm, the EXACT-U, to report utilities from the EXACT for use in cost-effectiveness studies for the UK.

II. METHODS

This study was conducted in five stages, using methods similar to the conversion of the SF-36 to the SF6D.[17] The first stage was to identify a subset of items and levels from the original instrument that maintains the factor structure and severity range for the preference-based instrument. The second stage was to design a set of EXACT-U health states to be valued using a preference elicitation technique. The third stage involves conducting a valuation survey with a sample of the UK general public. The fourth stage was to develop a range of econometric

models to predict utilities from the new classification system. The fifth stage was to test the developed model to evaluate the predictive, discriminant, and responsive properties. Methodologies for each step are detailed below.

A. Item Identification

The EXACT is a patient-reported daily diary developed to evaluate the frequency, severity, and duration of exacerbations of COPD and chronic bronchitis.[16] The instrument includes an assessment of respiratory and systemic attributes of an exacerbation, including chest discomfort (3 items), cough and sputum (2 items), and shortness of breath (5 items), followed difficulty with mucus, sleep disturbance, psychological state, and weak/tired (copyright restrictions prevent showing full measure). Change in total score on the EXACT is used to determine frequency, severity, and duration of events in clinical trials.

With 14 items and approximately five levels for each item, the EXACT is capable of providing estimates for over six billion health states. Previously collected EXACT patient data and input from clinical experts were used to identify a subset of items that would adequately cover the range of exacerbation severity for reporting utilities while decreasing the number of health states to be valued, reducing constraints on data collection and allowing for a more purposive coverage of health states valued.[17] A mixed-method approach involving classical test theory and item response theory with Rasch analysis was used in the item selection process.

The EXACT-U, for reporting utilities, consists of 5 items, covering chest discomfort, cough, shortness of breath with activity, psychological state, and weak/tired with 3 to 5 levels each ranging from not at all to severely/extremely (copyright prevents showing items or health states).

Given the focus of this paper to detail the methods for health state valuation and model development results, the item identification and content validation stages will be included in a separate manuscript.

B. Health State Development

Health states were assembled into one set for model development (Development Group) and a second set for model testing (Validation Group). This ensured one set of health states for the model development, and an independent second set of health states for the validation analyses. The performance of the model could then be tested against a separate set of health states valued by different respondents, thereby eliminating one source of bias.

Development Group health states were derived through three methods for maximum coverage and diversity: Identifying an orthogonal array of dimension-levels that have an equal chance of being combined with all levels of the other dimensions; analyzing previously collected EXACT data for commonly reported health states; and selecting corner states or one item at its worst level and all other items at their best level. The orthogonal array of health states was identified using a program in SPSS, which reported 24 health states for valuation. Evidence-based health states were analysed next by stratifying exacerbation severity and duration for a range of experiences. The evidence-based states were included to develop realistic scenarios likely to cover a range of severity.[17] Without patient data, health states are typically derived from clinician opinions and may not represent the actual experiences of patients.[18] The corner states were included to balance the mixed-level states with extreme states. The total number of 46 states was limited by resources to ensure an adequate number of observations by a set number of responders.

Validation Group health states were derived using the Rasch threshold map, selecting states by incremental level change in instrument severity (ie. 11111, 12111) (Figure 1). This method of health state selection has been previously used, with results suggesting there is a good match to patient-observed health states.[19] This method also provided a logical ordering of health states, which allowed for easier comparisons of predicted to actual health state utilities for identifying any inconsistencies in predicted values. The number of health states was limited by the Rasch output and the overlap with previously identified health states from the data. A total of 14 states were included.

Both groups had the best and worst states included for comparative purposes. A total of 60 unique health states were developed out of a possible 960, with 46 in the Development Group and 14 in the Validation Group.

C. Data Collection and Preparation

1. Data Collection

The instrument will aim to be used for UK health technology assessments, therefore methodology for the valuation study followed NICE recommendations where applicable (NICE 2008). The NICE guidance encourages condition-specific instruments used for reporting utilities to have development methods consistent with the EQ5D to provide comparability. The methods below were selected based on their consistency with the guidance and theoretical strength.

Sample size was based on the minimal number of valuations needed for each health state, due to limited funding available. From previous studies, approximately 50 observations per health states would allow adequate health state differentiation.[17, 20] A total of 400 UK general public members were sought, for an estimated sample of 350 for the Development Group, and 50 respondents for the Validation Group.

The study protocol was approved by LSHTM ethics committee prior to recruitment. Respondents were recruited from an existing database of those who had previously participated in TTO studies as well as from local universities, online Gumtree postings, and word of mouth.

At the start of the interview respondents completed a consent form, socio-demographic form including experience with lung problems, and the EQ5D. Health states were randomized within each group to ensure all states are valued an equal number of times and to avoid context biases from health state blocks.[21] Each respondent valued 11 randomized states from either the Development Group or Validation Group, including the best and worst, which were shuffled prior to each task.

A ranking exercise was conducted first, with all respondents arranging the health states in order from best to worst imaginable. The TTO exercise followed, using a time board prop, with respondents valuing each health state on a 10 year scale from "full health" to "dead". "Worse than dead" valuations were included by presenting some amount of time in the health state being valued, followed by the remaining time in full health.

The ping-pong method was used to reach their indifference point due to the recommendation that biases are minimized using this method.[22] Health states were traded down to the 3-month time frame in order to elicit the most precise utilities but to avoid respondent fatigue with shorter time frames. At completion of the interview, all respondents were offered £15 as a token of thanks for their time.

2. Data preparation

Utilities from the TTO exercise were transferred to fit a 1 to -1 utility range.[1] States rated as "better than dead" were calculated using $h_i=x/t$; where x is the amount of time in full health reached for that health state h_i, determined from the interview, and t is the life expectancy of 10 years. In order to anchor the states rated "worse than dead" to the -1 scale, the formula $h_i=-x/t$ was used. Time x is the full health endured for that health state h_i, and t is the life expectancy of 10 years.[23]

Respondents were examined for extreme responses and logical inconsistencies. Extreme responses are defined by respondents who have valued all health states utilities as 0.9875 or greater. Respondents who have valued all states with the same utility value, including the worst state, are deemed not to have understood the task at hand and will be removed from analyses.[17, 24, 25] Removing extreme responses allows for a more natural distribution of the utility values, as the lowest states are not inflated by the extreme values.

Logical inconsistencies occur when a logically less severe health state (ie. 11112) is rated as being worse than a logically more severe health state (ie. 11115).[26-29] For analyses of this study, the proposed method by Dolan and Kind[27] was followed, where health states were examined in pairs. An inconsistency existed when one health state with a less severe level on any attribute was rated as worse than another state with a more severe level on the same

attribute, given all other attribute levels were the same.[27] Numbers of pair-wise inconsistencies and the extent of differences were tallied and analysed for each valuation group. Respondents with inconsistent values greater than 0.10 and with more than 3 occurrences were removed.

Ideally the best practice is to include all respondents in the analyses; however extreme responses and logical inconsistencies have the potential to adversely affect the final model and its predictive ability. Consistent with previous methods, the most extreme and illogical respondents were evaluated prior to model development.[27] The decision to remove respondents was based on the extent of the responses and the effect on the developed models.

D. Model Development

Health states from the Development Group were used to derive a best-fitting regression model to accurately predict utilities for health states reported from the EXACT-U. This method was used in the development of the EQ5D, SF6D, and the construction of the OAB6D.[5, 21, 30] STATA version 10 was used for all data analyses.[31]

Models followed the common form:

$$y_{ij} = g\,(\beta^1 X_{ij} + \theta^1 R_{ij} + \delta^1 Z_j) + \varepsilon_j$$

The dependent variable, y_{ij}, is the TTO score for the health state i valued by respondent j. X is a vector of binary dummy variables. R term is a vector of terms to account for interactions between levels of states. Z are the personal characteristics that could affect an individual's valuation e.g. age, or sex.

Because the TTO valuations used "full health" as the upper anchor, rather than the best health state possible, it is not necessary to constrain the model with an intercept.[32] This allows the possibility of the best state to depart from one.

Model estimation began by employing the simplest application first in using Ordinary Least Squares (OLS) regression models. The first model specified included only the EXACT-U attributes. Subsequent models were built upon the original frame based on performance and

perceived model needs, including, but not limited to: combining attribute levels, addition of severity binary dummy variables, and demographic characteristics.

A series of mean models were developed next, to determine if a reduction in the data variance improves model fit. Mean models have the potential to average out the variance in the data, potentially reducing prediction error and including more significant coefficients in the model.

Last, either fixed or random effects models were developed. When using panel data, or multiple data points for each respondent, either of these models can be assumed to address the panel structure.[33] The hausman test was employed to determine which model should be used in this case. This test determines if the coefficients estimated by the efficient random effects estimator are the same as the ones estimated by the consistent fixed effects estimator. If they are (insignificant P-value, Prob>chi2 larger than .05), then it is appropriate to use random effects, otherwise the model produces biased estimators.[34]

When comparing models, inconsistencies in regression coefficients were of highest priority. Other considerations included the significance of coefficients and goodness of fit statistics using R^2. The models included demographics such as age, gender, and/or education level to explore their effects and determine the extent of inclusion with the model.

E. Model Validation

Model validation is an essential step in the development process. For this study, model validation was conducted in two stages, using two sets of data. First, the predictive ability was tested using the Validation Group data in order to select the best performing model. The error between predicted and actual utilities collected from respondents was calculated. Next, the selected model was tested for discriminant ability and responsiveness using patient data from the EXACT. Analyses include evaluating the ability of the EXACT-U to differentiate an exacerbation from stable state and recovery of an exacerbation through clinician-defined resolution, as well as to distinguish COPD exacerbation severity levels.

1. Predictive Properties

Models were evaluated on a number of parameters including root mean squared error (RMSE) and mean absolute error (MAE) between predicted and directly valued utilities. Graphs of the predicted versus actual utilities and the error versus actual utilities were plotted. The range of utility values from each model was reported.

All parameters on which the algorithms are evaluated will be aggregated. The algorithm with the greatest range of responses, least amount of error, and threshold limits were flagged for selection. Prioritisation was with the algorithm with the least number of logical health state utility inconsistencies.

2. Discriminant Validity and Responsiveness

The best performing model was selected for further testing using patient data, to examine the psychometric properties of the algorithm. The preferred model coefficients were applied to the EXACT diary data to report utilities for the stable group and for the exacerbation group from onset to clinician-diagnosed resolution. Discriminant validity and responsiveness were tested. Discriminant validity is the extent to which an instrument can differentiate among known groups.[35] Responsiveness is a component of construct validity and refers to the ability of an instrument to reflect underlying change, in this case response to treatment.[35]

EXACT patient data includes a sample size of 222 exacerbation patients over two months' duration and 189 stable patients over a seven day period. Clinical parameters include underlying COPD severity, exacerbation severity, and resolution based on clinician assessment.

The EXACT-U was analysed for the extent to which it can differentiate among stable COPD, a clinician-diagnosed exacerbation, and resolution. Testing was conducted on the extent to which the EXACT-U can differentiate among exacerbation severity levels. Differentiation was assessed by level of significance reached between tested groups. Data from Day 1 was used for discrimination of exacerbation severity levels, as well as between acute and stable states.

Responsiveness was evaluated by effect size using the standardized response mean (SRM). The SRM is defined as the mean difference between the baseline scores and the follow-up scores (i.e., mean change scores) divided by the standard deviation of the baseline scores. The size of

the SRM was then evaluated in terms of its meaningfulness.[35] An effect size of approximately 0.20 was considered small, one of 0.50 indicates moderate responsiveness, and an effect size ≥ 0.80 was considered highly sensitive. Responsiveness testing was conducted from Day 1 (Baseline) through to Day 14 or until clinician-diagnosed resolution has been indicated.

III. RESULTS

A. Data Collection

1. Respondent Characteristics

A total of 400 TTO interviews were completed in the UK, with 350 in the Development Group and 50 in the Validation Group. Two respondents in the Development Group were protocol failures and removed, leaving an analysis set of 398 respondents.

Demographic characteristics are presented in Table 1. The total sample was proportionally consistent in both the Development and Validation Groups. On average, the sample was white (71.1%), female (61.1%), 36 (13 SD) years old, with a university degree (53.8%), and employed full time (42%). Overall the sample reported good health, with EQ5D value of 0.94 (0.1 SD) and VAS of 82.1 (12.6 SD). Of the 28 respondents marked as having lung problems, 3 reported COPD, while the other 25 reported asthma. A total of 72 knew someone with lung problems, with asthma again being the most common diagnosis.

2. Health State Values

Of the 348 respondents in the Development group, 6 were extreme cases, or valued all health states as 0.9875 or above, and were removed from the sample for analysis.

With regards to inconsistencies, there were very few respondents seen as having logical inconsistencies. As a total, the highest number of any inconsistencies seen by a single respondent was 7 (n=1), 6 (n=2), 5 (n=3), and 4 (n=7). Only 11% of all opportunities to be inconsistent were taken by respondents, which results in more consistent group than in previous studies.[28] From these analyses, the decision was to keep all respondents in the analyses, however model performance was tested by removing the most inconsistent respondents for each developed model to confirm there was little impact on utilities or error.

Health state means ranged from 0.063 (worst state) to 0.978 (best state) with large standard deviations (0.290). States were mostly ordered properly for the group, with less severe states being rated with a higher utility than more severe states. Only four health states showed as being mis-ordered. Health states 11422, 22322, 32322, and 43432 had a mean utility value slightly higher than the next logically less severe health state. Overall the health states were skewed left, as expected, with the more severe states demonstrating a bimodal distribution.

B. Model Development

The results of developed models are presented in Table 2. All models were examined by significance of coefficients, number of inconsistencies, R^2, and range of utilities predicted (best state to worst state) from the algorithm before being analysed by the Validation group.

Models 1 through 4 were developed using OLS. To start, the model only included the EXACT-U attributes (Model 1). Coefficients for levels 1 and 2 from the cough (COU) and shortness of breath (SOB) attributes were positive, therefore in the opposite direction as expected. The levels were merged for both attributes for the next model (Model 2). An N_max binary dummy variable term was added to Model 3, which is similar to the N3 term in the EQ5D algorithm, to account for the additional decrement of any level being at its worst.[21] Age was added in Model 4 as a three-level term to evaluate the extent to which demographics influence the model fit.

Based on the performance of the above models, a series of mean models were developed. Mean utility values of all the health states were taken and used to develop the mean models. Again, only the EXACT-U attributes were included first (Model 5). The coefficient for the COU attribute level 1 was positive, which suggested the merging of levels 1 and 2 for Model 6. The N_max term was significant in the previous models, and therefore was added to the mean model to improve fit (Model 7). Mean data prevented the inclusion of the age term for these models.

The third round of models developed planned to use either random or fixed effect designs. Results of the hausman test (Prob>chi2 =0.0005) determined that fixed effects models should be used. Model 8 included only EXACT-U attributes and all levels. Again, the direction of

coefficients for the COU attribute for level 1 was positive and suggested levels 1 and 2 be combined for Model 9. The N_max term was added next to account for the extra decrement in the severe levels, based on the strong performance in earlier models (Model 10).

All models that included levels 1 and 2 for COU predicted inconsistent utilities. This justified removal of Models 1, 5, and 8 from further selection. As expected, the R^2 reported from the individual data models was nearly 30% of that from the mean models, at 0.32 compared to 0.91. The mean models eliminate the individual variance for health state values. This, in addition to the lower MAE, resulted in the mean models being slightly favored over the individual models. Models 6, 9, and 10 then remained for the next step, predictive ability testing.

C. Model Validation

1. Predictive Properties

The Validation group was used to examine predicted utilities from the algorithm compared to actual utility values collected. The models were tested for prediction error by MAE and RMSE. Final model selection was based on the collection of evidence supporting the best performance. The MAE and RMSE are slightly lower with the mean models than the individual models, which also supports the decision for elimination in favor of the mean models. Of the two mean models remaining, the N_max term in Model 7 is non-significant, with all other error statistics remaining the same, which supports the selection of Model 6, without the N_max term, from the more simple mean design.

The fixed effects (FE) models do not report an R^2, so they were not compared in this context. The MAEs for the FE models were consistent with earlier models, however the RMSE was at least 0.146 lower with the FE models than any other model, particularly Model 6. The spread of utilities from the mean models appears to be slightly larger than the FE models. However, the FE models had more coefficients reaching significance levels, and with the lower RMSE, it appears to be a better selection of models. Of the two remaining models (Model 9 and 10), Model 10 was selected due to the statistical significance of the N_max term, the greater number of significant coefficients, and the lower RMSE. Model 10 will be used for further

validity testing. Figure 2 shows the Model 10 predicted utilities against the actual utilities for the Validation group.

2. Discriminant Validity and Responsiveness

The algorithm from Model 10 was derived by applying all coefficients to the levels and attributes of the EXACT-U. The Model 10 algorithm was applied to the patient data to evaluate discriminant validity and responsiveness.

Discriminant testing began with analyses between stable and acute patients. Utilities were 0.780 for stable and 0.631 for acute and reached statistical significance ($p<0.001$), supporting the EXACT-U differentiation between stable COPD and onset of an exacerbation. Next the algorithm was tested for differentiation among the exacerbation severity levels, from mild to very severe, using the clinician assessment of severity. Utilities declined from mild to very severe, with statistical significance ($p=0.01$ for mild/moderate; $p<0.001$ for moderate/severe) reached for all levels with the exception of severe to very severe. Despite a difference in utilities, 0.574 for severe exacerbations compared to 0.536 for very severe, the lack of statistical significance could be due to the small sample size in the very severe group ($n=4$) compared to the severe group ($n=33$). Figure 3 shows the spread of utilities for stable COPD and each level of exacerbation severity, as diagnosed by a clinician.

Responsiveness was tested using standardized response mean (SRM). SRM for the EXACT-U at Day 3, Day 7, Day 10, and Day 13 was, 0.38, 0.58, 0.68, and 0.77 respectively. SRM of 0.50 and greater suggests good responsiveness to daily change due to intervention during an exacerbation. The responsiveness seen at Day 3 suggests the EXACT-U is able to report sensitive utilities immediately after treatment has begun and effects are noted.

Patients who experience an exacerbation were shown to have, on average, a lower utility at clinician-diagnosed resolution of an exacerbation compared to the stable patients. This is consistent with previous research on exacerbations.[36]

IV. DISCUSSION

Responses on the EXACT correspond to items and levels selected for the EXACT-U, which allows data from the daily diary to be used in economic evaluations. Predictive testing suggests minimal error between predicted and directly measured utilities. Discriminant testing resulted in differentiation between acute and stable states, and exacerbation severity levels from mild through severe. Responsiveness of the EXACT-U is strong as well, where change associated with treatment was detected after the first few days of an exacerbation diagnosis and through at least the first two weeks, where most change occurs. Patients who have a clinician-confirmed resolution from an exacerbation, at least 4 weeks post, are seen to have lower utilities as a group, than the patients who were stable.

The EQ5D is recommended for health status measurement due to the generic application of the dimensions to a range of disease areas.[2] Poor discriminant validity and responsiveness with regards to COPD have been documented in numerous studies.[1, 10, 11, 37, 38] One study showed the EQ5D could differentiate the acute phase of a moderate exacerbation from the recovery phase in patients from day 1 to day 14 or 42, but not change from day 1 to day 8.[37] By contrast, Seemungal [39]reported that the greatest change in symptoms occurs within the first week of an exacerbation. There has also been difficulty in comparing utility values of the EQ5D across studies within COPD, presumably due to differences in inclusion/exclusion criteria and co-morbidities. As reported in one study, the absolute number of concomitant diagnoses was significantly associated with worse EQ-5D utility scores ($p<0.001$), independent of GOLD stage, while there was no significant interaction between GOLD stages and the presence or absence of co-morbidity.[40]

In addition to concerns regarding the EQ5D, inconsistent modeling trends further misrepresent exacerbations due to the lack of patient-reported exacerbation utilities. Current trends to model exacerbations use either a percentage decrement from a stable COPD utility, a single utility to account for a range of exacerbation severity levels, or two utility values to account for a moderate and severe exacerbation.[10, 41-44] The duration for exacerbation states can be randomly assigned based on severity, for the duration of the cycle at weeks to months, or left

unreported. While this may be acceptable for evaluations of interventions to treat stable COPD, interventions that impact the frequency, duration, or severity of an exacerbation need more accurate evidence accounting for treatment benefits. Evidence from the EXACT-U suggests exacerbation severity levels do have significantly different utilities. Patient reported EXACT-U utilities could reflect the rapid change in health over the course of an exacerbation, the average duration of an exacerbation and the impact of longer lasting exacerbations, or the total change in utility from the start of an exacerbation through resolution.

EXACT-U utilities can be used in conjunction with the EQ5D or alone. The HRQL change during an exacerbation can be reflected in the EXACT-U, while the generic instrument can report stable COPD states. The benefit to this coordination is the use of the EQ5D for COPD, consistent with reimbursement agencies' recommendations, but more precise measurement of exacerbations, given the limitations of the EQ5D in that area. The limitation of this approach is the combination of generic and condition-specific utility values together. The instruments measure different constructs and components of a patient's HRQL, and would therefore reflect change to potentially different degrees. Combining both, to use an EQ5D utility for stable COPD and employ the EXACT-U to show decrements for exacerbation states, may inaccurately reflect health states.

Preliminary research suggests it may also be able to report utilities for the range of COPD severity as well for exacerbations. This would eliminate any harmonization concerns, and cover all COPD utility needs with one instrument. Further testing on a stable COPD dataset would be necessary in this care prior to use.

The difficulty in using these measures to report utilities for economic evaluations arises when comparing across disease areas for resource allocation. While utilities measured from different measures tend to prevent direct comparisons, it could be more appropriate to adequately capture intervention impact within disease areas, and then compare QALYs across diseases.

Limitations of this study include the small sample size of the health state valuation study and the limited data available to test the algorithm. At present time, results cannot be generalized

to other disease areas. Caution should be used when using exacerbation utilities as testing has only been conducted on one dataset.

V. CONCLUSIONS

The EXACT-U is a condition-specific preference-based measure designed using methods similar to the EQ5D development. Utilities reported from the instrument can be used where EQ5D data are not available or where condition-specific utilities are preferred to report the impact of the severity, duration, and frequency of exacerbations on patients. The instrument can be used in cost-effectiveness assessments to report utilities for exacerbating COPD patients from clinical trials where EXACT patient data are available.

The utility instrument shows strong discriminant properties among exacerbation severity levels and responsiveness to exacerbation change and resolution. Ongoing analyses are being conducted to improve the final model and ascertain additional reliability properties.

Acknowledgements

We would like to thank Forest Research Institute for providing an unrestricted grant to aid data collection, and to United BioSource Corporation for providing access to internal resources for the study, particularly E Davies and S Macker for database development and data entry review. We would also like to thank W Wedzicha, J Quint, and J Hurst at Hampstead Hospital for recruitment of patients for EXACT-U health state content validation. Last, this study could not have been conducted without the support and expertise of N Leidy and P Jones.

VI. REFERENCES

1. Brazier, J., et al., *A review of the use of health status measures in economic evaluation.* Health Technology Assessment, 1999. **3**(9): p. iii-163.
2. (NICE), N.I.f.C.E., *Technical guidance for manufacturers and sponsors on making a submission to a technology appraisal*. 2008: UK.
3. Haywood, K.L., et al., *EuroQol EQ-5D and condition-specific measures of health outcome in women with urinary incontinence: reliability, validity and responsiveness.* Qual Life Res, 2008. **17**(3): p. 475-83.
4. Chang, S.W., et al., *The impact of diuretic therapy on reported sexual function.* Arch Intern Med, 1991. **151**(12): p. 2402-8.
5. Yang, Y., et al., *Estimating a preference-based single index from the Overactive Bladder Questionnaire.* Value Health, 2009. **12**(1): p. 159-66.

6. Yang, M., *Estimating a preference-based single index from the Asthma Quality of Life Questionnaire (AQLQ)*, in *Health Economics and Decision Science*. 2007, University of Scheffield, School of Health and Related Research.
7. Espallargues, M., et al., *The impact of age-related macular degeneration on health status utility values.* Investigative Ophthalmology and Visual Science, 2005. **46**(11): p. 4016-4023.
8. Tandon, P.K., H. Stander, and R.P. Schwarz, Jr., *Analysis of quality of life data from a randomized, placebo-controlled heart-failure trial.* J Clin Epidemiol, 1989. **42**(10): p. 955-62.
9. Stahl, E., et al., *Health-related quality of life is related to COPD disease severity.* Health Qual Life Outcomes, 2005. **3**: p. 56.
10. Borg, S., et al., *A computer simulation model of the natural history and economic impact of chronic obstructive pulmonary disease.* Value Health, 2004. **7**(2): p. 153-67.
11. Szende, A., et al., *Estimating health utilities in patients with asthma and COPD: evidence on the performance of EQ-5D and SF-6D.* Qual Life Res, 2009. **18**(2): p. 267-72.
12. O'Reilly, J.F., A.E. Williams, and L. Rice, *Health status impairment and costs associated with COPD exacerbation managed in hospital.* Int J Clin Pract, 2007. **61**(7): p. 1112-20.
13. Chiou, C.F., et al., *Development of the multi-attribute pediatric asthma health outcome measure (PAHOM).* International Journal for Quality in Health Care, 2005. **17**(1): p. 23-30.
14. Torrance, G.W., et al., *Development and initial validation of a new preference-based disease-specific health-related quality of life instrument for erectile function.* Quality of Life Research, 2004. **13**(2): p. 349-359.
15. Revicki, D.A., et al., *Integrating patient preferences into health outcomes assessment: The multiattribute asthma symptom utility index.* Chest, 1998. **114**(4): p. 998-1007.
16. Leidy, N.K., Powers, J.H., Howard, K.A., Petrillo, J.M., Wilcox, T.W. and EXACT-PRO Study Group, *The EXACT-PRO Initiative: Development of a Standardized Outcome Measure for Evaluating Exacerbations of Chronic Obstructive Pulmonary Disease*, in *American Thoracic Society International Conference*. 2008: Toronto, Ontario, Canada.
17. Brazier, J., J. Roberts, and M. Deverill, *The estimation of a preference-based measure of health from the SF-36.* Journal of Health Economics, 2002. **21**(2): p. 271-292.
18. Chancellor, J.V.M., D. Coyle, and M.F. Drummond, *Constructing health state preference values from descriptive quality of life outcomes: Mission impossible?* Quality of Life Research, 1997. **6**(2): p. 159-168.
19. Young, T., et al., *The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis.* Qual Life Res, 2009. **18**(2): p. 253-65.
20. Stevens, K., et al., *Multi-attribute utility function or statistical inference models: A comparison of health state valuation models using the HUI2 health state classification system.* Journal of Health Economics, 2007. **26**(5): p. 992-1002.
21. Kind, P. (1996) *The EuroQoL Instrument: An index of health-related quality of life*. Quality of Life and Pharmacoeconomics in Clinical Trials **Volume**,
22. Bennett, K.J., et al., *Cost-utility analysis in depression: The McSad utility measure for depression health states.* Psychiatric Services, 2000. **51**(9): p. 1171-1176.
23. Stalmeier, P.F.M., et al., *The gap effect: Discontinuities of preferences around dead.* Health Economics, 2005. **14**(7): p. 679-685.
24. Kaplan, R.M. and J.P. Anderson, *A general health policy model: update and applications.* Health Serv Res, 1988. **23**(2): p. 203-35.
25. Torrance, G.W., et al., *Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2.* Med Care, 1996. **34**(7): p. 702-22.
26. Badia, X., M. Roset, and M. Herdman, *Inconsistent responses in three preference-elicitation methods for health states.* Soc Sci Med, 1999. **49**(7): p. 943-50.
27. Dolan, P., et al., *Valuing health states: A comparison of methods.* Journal of Health Economics, 1996. **15**(2): p. 209-231.
28. Lamers, L.M., et al., *Inconsistencies in TTO and VAS values for EQ-5D health states.* Medical Decision Making, 2006. **26**(2): p. 173-181.
29. Devlin, N. and D. Parkin, *Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis.* Health Economics, 2004. **13**(5): p. 437-452.

30. Brazier, J., et al., *Deriving a preference-based single index from the UK SF-36 Health Survey.* Journal of Clinical Epidemiology, 1998. **51**(11): p. 1115-1128.
31. StataCorp, L., *STATA 10*. 2007.
32. Brazier, J.R., J; Salomon, J; Tsuchiya, A, *Measuring and Valuing Health Benefits for Economic Evaluation*. 2007, Oxford: Oxford University Press.
33. Hardin, J.H., J, *Generalized Linear Models and Extensions*. Second ed. 2007, College Station, Texas: Stata Press.
34. Stock, J.W., M *Introduction to Econometrics*. Second ed. 2003: Addison-Wesley.
35. Nunnally J, B.I., ed. *Psychometric Theory*. 3rd ed. 1994, McGraw Hill: New York.
36. Spencer, S. and P.W. Jones, *Time course of recovery of health status following an infective exacerbation of chronic bronchitis.* Thorax, 2003. **58**(7): p. 589-93.
37. Goossens, L., M. Rutten-van Molken, and C. Nivens, *IS THE EQ-5D RESPONSIVE TO RECOVERY FROM A MODERATE COPD EXACERBATION?*, in *ISPOR 11th Annual European Conference*. 2008: Athens, Greece.
38. Singh, S.J., et al., *A comparison of three disease-specific and two generic health-status measures to evaluate the outcome of pulmonary rehabilitation in COPD.* Respir Med, 2001. **95**(1): p. 71-7.
39. Seemungal, T.A., et al., *Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease.* Am J Respir Crit Care Med, 2000. **161**(5): p. 1608-13.
40. Rutten-van Molken, M.P., et al., *Does quality of life of COPD patients as measured by the generic EuroQol five-dimension questionnaire differentiate between COPD severity stages?* Chest, 2006. **130**(4): p. 1117-28.
41. Sin, D.D., K. Golmohammadi, and P. Jacobs, *Cost-effectiveness of inhaled corticosteroids for chronic obstructive pulmonary disease according to disease severity.* Am J Med, 2004. **116**(5): p. 325-31.
42. Hoogendoorn, M., et al., *A dynamic population model of disease progression in COPD.* Eur Respir J, 2005. **26**(2): p. 223-33.
43. Oostenbrink, J.B., et al., *Probabilistic Markov model to assess the cost-effectiveness of bronchodilator therapy in COPD patients in different countries.* Value Health, 2005. **8**(1): p. 32-46.
44. Spencer, M., et al., *Development of an economic model to assess the cost effectiveness of treatment interventions for chronic obstructive pulmonary disease.* PharmacoEconomics, 2005. **23**(6): p. 619-637.

Figure 1: Rasch Threshold Map

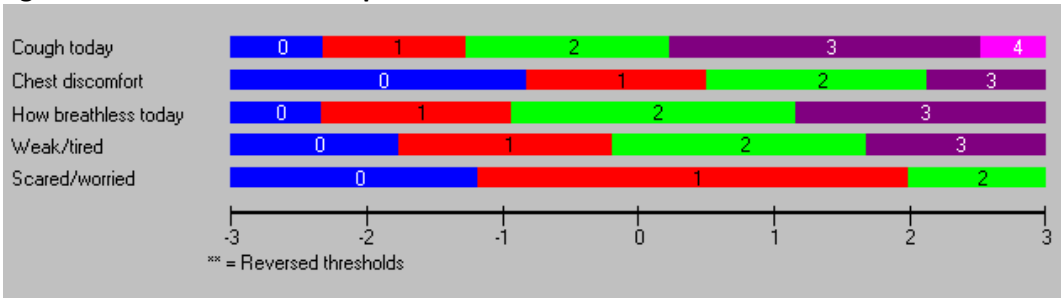

Figure 2: Model 10 Predicted vs Actual Utilities Graph

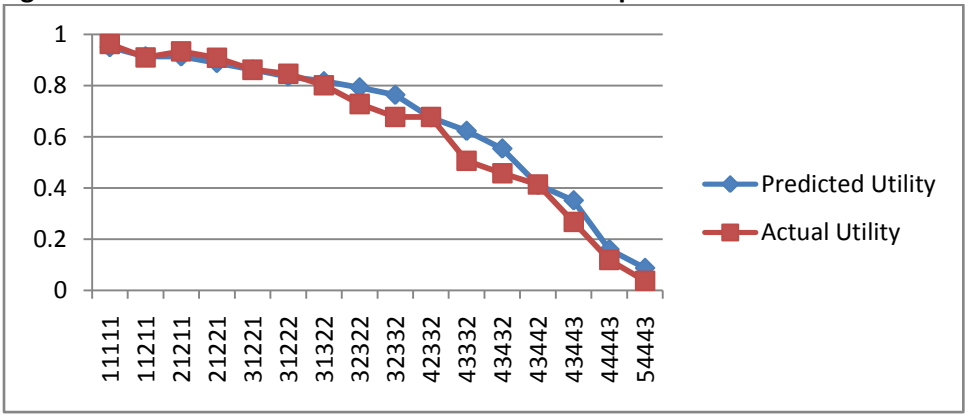

Figure3: Model 10 Predicted Utility for Stable COPD and Exacerbations by Severity

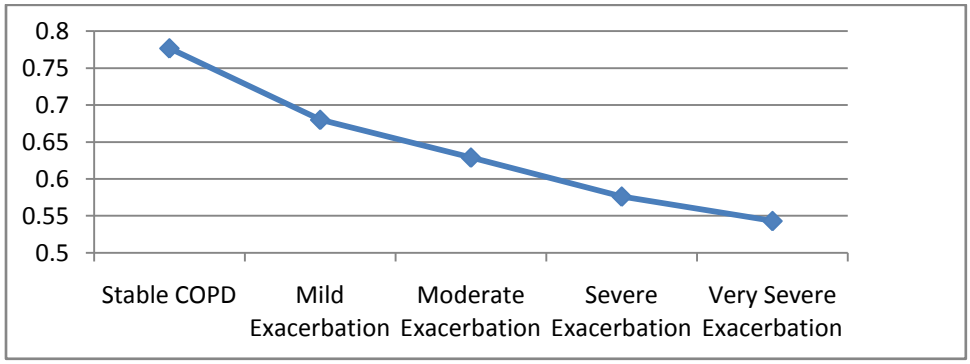

Table 1: Demographic Characteristics

	Develop (n=348)	Valid (n=50)	Total (N=398)
Age, mean (SD)	35.8 (13.1)	36.8 (15.4)	35.9 (13.4)
Gender, % male	38.22	44	38.94
Ethnicity, N(%)			
White (British, Irish, Other)	245 (70.4)	38 (76.0)	283 (71.1)
Black (Caribbean, African, Other)	52 (15.0)	5 (10.0)	57 (14.3)
Middle Eastern (Indian, Pakistan, Bangladeshi)	24 (6.9)	3 (6.0)	27 (6.9)
Asian (Chinese, Other)	17 (5.0)	2 (4.0)	19 (4.9)
Mixed	17 (5.0)	3 (6.0)	20 (5.1)
Education, N(%)			
No Formal Education	5 (1.4)	0 (0)	5 (1.3)
GCSE or High School	31 (8.9)	7 (14.0)	38 (9.6)
A-levels	44 (12.6)	11 (22.0)	55 (13.8)
Vocational	34 (9.8)	4 (8.0)	38 (9.6)
University Degree	191 (54.9)	23 (46.0)	214 (53.8)
Post-Graduate	25 (7.2)	3 (6.0)	28 (7.0)
Other	18 (5.2)	2 (4.0)	20 (5.0)
Employment, N(%)			
Full-time	150 (43.1)	17 (34.0)	167 (42.0)
Part-time	51 (14.7)	8 (16.0)	59 (14.8)
Homemker	10 (2.9)	2 (4.0)	12 (3.0)
Student	77 (22.1)	13 (26.0)	90 (22.6)
Unemployed	28 (8.0)	3 (6.0)	31 (7.8)
Retired/Disabled	20 (5.7)	4 (8.0)	24 (6.0)
Self-Employed	11 (3.2)	3 (6.0)	14 (3.5)
Other	1 (0.3)	0 (0)	1 (0.3)
EQ5D, mean (SD)	0.95 (0.1)	0.95 (0.1)	0.94 (0.1)
EQ5D vas, mean (SD)	82.1 (12.6)	81.7 (13.0)	82.1 (12.6)

Table 2: EXACT-U Developed Models and Performance

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**	**Model 10**
	OLS	OLS COU/SOB12	OLS COU/SOB12N	OLS COU/ SOB12N A	OLS M	OLS MCOU12	OLS MCOU12N	FE	FE COU12	FECOU12 N
_ICOU_2	0.051	X	X	X	0.049	X	X	0.041	X	X
_ICOU_3	0.020	-0.005	-0.007	-0.008	0.022	-0.007	-0.008	-0.005	-0.025	-0.026
_ICOU_4	-0.062	-0.085	-0.083	-0.089	-0.092	-0.119	-0.121	-0.098	-0.118	-0.115
_ICOU_5	-0.160	-0.174	-0.167	-0.169	-0.240	-0.261	-0.254	-0.182	-0.194	-0.188
_ICHS_2	-0.047	-0.040	-0.044	-0.042	-0.049	-0.046	-0.046	-0.025	-0.021	-0.024
_ICHS_3	-0.049	-0.043	-0.041	-0.042	-0.087	-0.083	-0.083	-0.082	-0.079	-0.076
_ICHS_4	-0.272	-0.270	-0.263	-0.263	-0.252	-0.251	-0.247	-0.273	-0.273	-0.267
_ISOB_2	0.014	X	X	X	-0.006	-0.004	-0.006	-0.032	-0.029	-0.036
_ISOB_3	-0.038	-0.042	-0.046	-0.044	-0.057	-0.056	-0.059	-0.051	-0.048	-0.054
_ISOB_4	-0.091	-0.097	-0.082	-0.081	-0.094	-0.095	-0.090	-0.099	-0.099	-0.089
_IWT_2	-0.046	-0.031	-0.038	-0.037	-0.043	-0.035	-0.035	-0.032	-0.021	-0.026
_IWT_3	-0.062	-0.054	-0.058	-0.059	-0.071	-0.068	-0.067	-0.058	-0.052	-0.055
_IWT_4	-0.195	-0.192	-0.186	-0.188	-0.198	-0.196	-0.192	-0.204	-0.202	-0.197
_ISW_2	-0.017	-0.010	-0.012	-0.014	-0.026	-0.025	-0.024	-0.030	-0.027	-0.027
_ISW_3	-0.093	-0.091	-0.084	-0.087	-0.114	-0.115	-0.113	-0.093	-0.093	-0.087
_lage_cat_2				0.049						
_lage_cat_3				0.041						
N_max			-0.043	-0.040			-0.009			-0.034
_cons	0.900	0.911	0.931	0.906	0.916	0.936	0.939	0.925	0.933	0.950
N	3684	3684	3684	3684	49	49	49	3684	3684	3684
Inconsistent	3	0	0	0	2	0	0	1	0	0
Signif coeff	13	11	13	15	9	9	9	12	11	13
R2	0.322	0.321	0.322	0.325	0.913	0.906	0.906	n/a	n/a	n/a
MAE	0.234	0.236	0.242	0.235	0.232	0.232	0.231	0.234	0.235	0.233
RMSE	0.344	0.344	0.344	0.347	0.336	0.336	0.336	0.194	0.190	0.210
Pred Range	0.90 to 0.089	0.911 to 0.087	0.931 to 0.105	0.906 to 0.063	0.916 to 0.018	0.936 to 0.019	0.962 to 0.034	0.934 to 0.075	0.929 to 0.075	0.949 to 0.088