making, reduced patient benefits and less efficient use of public resources. Within individual agencies, prioritisation is also a business function that must balance the need to plan for and manage the allocation of resources with the need to provide expeditious advice to decision makers and adapt quickly to changing circumstances. This research describes a transparent and responsive framework for selecting health technologies to assess, minimising the potential for important technologies to be missed and providing a useful resource for HTA agencies facing similar issues. Topics are identified through a mix of routine horizon scanning, a formally convened advisory group consisting of the major decision makers from within the publicly funded health system and informal business intelligence gathering. Screening is carried out to eliminate technologies that are clearly unsuitable and provisionally grade all remaining candidates according to three principal criteria; 1) clinical impact (patient population, potential incremental effect and availability of alternatives); 2) economic impact (incremental costs and potential disruptive effect on how services are currently organised) and 3) policy impact (link to decision-making and factors that make it likely to feature on the national health care agenda). The screening process feeds into an in-depth expert group discussion, which also considers operational issues such as the extent of the advice required to inform the decision, data availability and costs associated with the assessment. We also describe a software visualisation tool developed to facilitate the prioritisation process, as well as measures for quality assurance and ongoing performance evaluation.

PRM239

GOAL ATTAINMENT SCALING – A USEFUL INDIVIDUALIZED CLINICAL OUTCOME MEASURE

Jones M, Kharawala S, Langham J, Gandhi P

Bridge Medical, London, UK

Goal Attainment Scales (GAS) capture outcomes relevant to individual patients and provide "real-world" outcome measurement. This abstract will describe the background to their use, their operationalization, strengths and limitations. Traditional outcome measures assess a standardised set of questions regardless of their relevance to each patient. GAS overcomes these weaknesses because it is an individualised assessment based on achievement of goals **which are personal** to each patient. Despite its widely cited use in academic literature and good psychometric properties it is rarely used in drug intervention studies. Operationalisation of GAS varies but follows these basic steps: 1) The patient's specific problem areas are assessed and goals for each defined; 2) A GAS for each goal is created and an "expected outcome" for each agreed; 3) Goal attainment levels are defined for each point on, typically, a 5-point scale (expected outcomes are usually scored 0; baseline is often -2, but may be -1 or 0 depending on potential for deterioration). Each level must be carefully described in a way that is relevant, observable, measurable and consistent with study design. Published standardised goals are available; and 4) A standardized statistical formula provides overall goal attainment. Benefits include: ease of use; relevant goals; no redundant items; assessment of multiple domains; provides quantifiable and applicable outcomes across different conditions and severities; potentially more sensitive measure than traditional scales. Limitations include: potential bias; appropriate goal selection and outcome prediction; observable changes may differ from pre-defined outcomes; time consuming; may require independent GAS assessors for blinded trials; may require "control" goals not affected by treatment; statistical issues around single overall score. In capturing those outcomes relevant to each individual patient, GAS has potential use in supporting product labeling claims and value assessment of a medicine by HTA and payers.

PRM240

AVOIDING AND IDENTIFYING ERRORS AND OTHER THREATS TO THE CREDIBILITY OF HEALTH ECONOMIC MODELS

Tappenden P, Chilcott J

University of Sheffield, Sheffield, UK

Health economic models have become the primary vehicle for undertaking economic evaluation and are used in various health care jurisdictions across the world to inform decisions about the use of new and existing health technologies. Models are required because a single source of evidence, such as a randomised controlled trial, is rarely sufficient to provide all relevant information about the expected costs and health consequences of all competing decision alternatives. Whilst models are used to synthesise all relevant evidence, they also contain assumptions, abstractions and simplifications. By their very nature, all models are therefore "wrong." Whilst the presence of imperfect evidence provides the impetus for developing models, it is also the reason why we can never fully validate them. As such, the interpretation of estimates of the cost-effectiveness of health technologies requires careful judgements about the degree of confidence that can be placed in the models from which they are drawn. The presence of a single error or inappropriate judgement within a model may lead to inappropriate decisions, an inefficient allocation of health care resources and ultimately suboptimal outcomes for patients. This study sets out a taxonomy of threats to the credibility of health economic models. The taxonomy segregates threats to model credibility into three broad categories (1) unequivocal errors, (2) violations and (3) matters of judgement, and maps these across the main elements of the model development process. These three categories of threats to model credibility are defined according to the existence of criteria for judging correctness, the degree of force with which such criteria can be applied, and the means by which potential threats can be handled. A range of suggested processes and techniques for avoiding and identifying these threats is put forward with the intention of prospectively increasing the credibility of any given model.

PRM241

ASSESSING HETEROGENEITY OF TREATMENT EFFECT USING REAL WORLD DATA

Murray JF[1], Kadziola Z[2], Zagar A[1]

[1]*Eli Lilly and Company, Indianapolis, IN, USA,* [2]*Eli Lilly Regional Operations GmbH, Vienna, Austria*

There is increasing scrutiny of pharmaceuticals on their value proposition as well as a growing demand for evidence on real world effectiveness once they are commercially available. There are many challenges in producing valid and reliable estimates of real world effectiveness. A major challenge is assessing a product's effectiveness relative to why patients may respond differently to a treatment (i.e., identifying groups of patients exhibiting "Heterogeneity of Treatment Effect" (HTE) using subgroup identification methods). Assessing HTE is critical to understanding differences that may exist between the efficacy observed in randomized clinical trials and a product's real world effectiveness. Understanding causes for HTE is required for correct attribution of any observed difference between efficacy and effectiveness to the product versus other sources (e.g., patient behavior); Not recognizing and accounting for HTE will confound assessment of a product's performance, which ultimately affects its acceptance and use by payers, physicians, and patients. Failure to define and incorporate subgroups is a frequent criticism of systematic evidence reviews and comparative effectiveness research reports. However, the analytical methods for finding factors that define subgroups that explain HTE are challenging due to many known statistical issues (e.g., limited statistical power, multiplicity adjustments) Real world data exacerbates the analytical challenges due in part to biases (e.g., selection bias) and issues (e.g., data quality) inherent in the data. We will describe the data and bias challenges that create these analytical complexities for detecting the cause and magnitude of HTE when using real world data. We will present results from a simulation experiment that compared and validated several subgroup methods developed to address these data and analytical issues. We simulated 22 permutations of subgroups with known identification criteria and treatment effects to determine the performance of the methods.

PRM242

IMPACTS OF EPRO DATA COLLECTION MODE SELECTION ON PATIENT INCLUSION

Holzbaur E, Ross J, Wade M, Rothrock T

Almac Clinical Technologies, Souderton, PA, USA

OBJECTIVES: The Electronic Patient Reported Outcome (ePRO) data collection mode selected for trials is often based on efforts to minimize timelines, budgets, and patient burden. However, are sponsors inadvertently introducing bias into trial results in this selection process? This conceptual paper reviews common ePRO modes and explores patient groups that may be excluded. **METHODS:** Common modes for ePRO data collection are reviewed. An assessment of potential patient groups that may be excluded is performed based on ePRO mode. **RESULTS:** Common ePRO modes include telephone, web, and handheld device. As sponsors look to reduce costs, improve data quality, and reduce patient burden, industry has continued its shift towards patients using their own telephone, computer, tablet, or smartphone and away from sponsors provisioning these devices to patients. Choice of patient-provisioned device: Patients from certain geographic areas may be excluded where internet connections and cellular/mobile telephone reception is limited. Requiring patients to use their personal web/mobile device may exclude patient groups with certain economic, cultural, or demographic characteristics who live in rural or underdeveloped areas. Choice of sponsor-provisioned device: Logistics issues, i.e. shipment of devices including customs considerations, reliability of data transmission, storage and replacement of devices and cords, training, etc. **CONCLUSIONS:** The objective of clinical trials is to establish treatment effectiveness, generalizable to the overall patient population. ePRO mode selection may impact inclusion of individuals from certain economic, cultural, demographic, and geographic areas. Exclusion of these groups could impact results; therefore, it is important to understand the potential bias that can be introduced when selecting an ePRO mode. Proper planning should include assessment of patient population and inclusion of regions that would render generalizability. ePRO mode selection should be based on which method works best for the required regions to optimize inclusion, as well as the patient population's characteristics to minimize burden.

PRM243

CLINICAL OUTCOME ASSESSMENT (COA) INSTRUMENT SCORING: THE VALIDITY AND PRECISION OF UNWEIGHTED SUMMARY SCORES VERSUS IRT WEIGHTED SCORES, AND THE ADDED VALUE OF IRT STANDARD ERRORS

Coon CD[1], Lenderking WR[2]

[1]*Adelphi Values, Boston, MA, USA,* [2]*Evidera, Lexington, MA, USA*

COA development experts in recent years have given thought to the psychometric evaluation of instruments and their ability to detect meaningful differences between patient groups. The scoring of the instruments, however, has received less attention, with various approaches sometimes suggested without a clear preference or justification. The score is ultimately used for evaluating patient outcomes and treatment efficacy and is what requires validation, so this seems like a significant omission. We examine the traditionally accepted unweighted summary score approach and compare it to the more complex IRT weighted scoring to evaluate if the gain in precision justifies the increased scoring complexity. Precision may differ depending on whether the score is close to the mean of the population or closer to the extreme ends of the distribution. Simulated data are used for this comparison to evaluate if the precision of the scores differs depending on the location of the score and if the instrument is used for group comparisons versus individual diagnosis. Additionally, we recognize that the reliability of a scale is likely to be variable across the range of its scores. With that in mind, we consider an approach to comparing mean scores between groups that incorporates the standard error of each individual IRT score into the model. By using the IRT standard errors, we can adjust for the different levels of uncertainty associated with ranges of scores along the scale, ultimately providing us greater confidence in the group comparison results.

PRM244

EVALUATION OF ESTIMATORS OF TREATMENT EFFECT IN OBSERVATIONAL STUDIES

Faries DE[1], Lipkovich I[2], Kadziola Z[3]