PROMs in clinical research
Tips for study design, implementation, analysis and interpretation

Professor Madeleine King
Quality of Life Office
PoCoG, University of Sydney
Monash University 25 June 2015

Learning Objectives

By the end of the workshop, participants will:
1. Understand the terms “quality of life” and “patient-reported outcome” (PRO), their components and the relationships among them
2. Appreciate the value of PROs as outcome measures in clinical research
3. Understand the principles of good study design for PRO endpoints
4. Be aware of the problems caused by missing PRO data
5. Be aware of the implementation strategies to minimise rates of missing PRO data
6. Be aware of sample size considerations for PROs
7. Appreciate the challenges in analysing and interpreting PRO data, and methods to address them
8. Be aware of key items that should be included in a study protocol about PRO endpoints, and the Univ Sydney QoL Office PROtocol checklist.
9. Be aware of the CONSORT-PRO reporting standards

How does disease affect a patient?

Disease + TMT

Proximal Effects

Disease
Symptoms:
- e.g. pain, fatigue, breathlessness
Treatment effects:
- Negative: toxicity
e.g. nausea, rash
- Positive:
- palliation e.g. pain relief, other

Distal Effects

Functioning & well-being
- Physical
- Emotional
- Social
- Role
- Cognitive
- Sexual, body image
- Spirituality
- Financial

Psychological
- Fear, anger, uncertainty, anxiety, depression

Process of care
- Satisfaction with health care/providers/information
- Preferences
- Inconvenience

Global QoL
- Well Being & Happiness

Other aspects of life
- Finances, family, work, leisure
- Spiritual

Terminology – can be confusing!

Broad umbrella: ‘Quality of Life’

‘Health-related quality of life (HRQoL)’

‘Patient-reported outcomes PRO’

“A measurement based on a report that comes directly from the patient about the status of a patient’s health condition without amendment or interpretation of the patient’s response by a clinician or anyone else.”

FDA Guidance (2009)
Why include Quality of Life (QoL) and/or related patient reported outcomes (PROs)?

**Intended use**

<table>
<thead>
<tr>
<th>How PROs can value-add</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Choosing the best treatment</strong></td>
</tr>
<tr>
<td>When PRO is primary outcome:</td>
</tr>
<tr>
<td>- palliative settings</td>
</tr>
<tr>
<td>- survivorship studies (e.g., impact of chronic side-effects)</td>
</tr>
<tr>
<td>When PRO is secondary outcome:</td>
</tr>
<tr>
<td>- may support the primary outcome (survival/recurrence)</td>
</tr>
<tr>
<td>- may counterbalance – trade-offs</td>
</tr>
<tr>
<td>- may complement clinical parameters in differentiating treatments with equivalent survival outcomes</td>
</tr>
<tr>
<td><strong>Enriching understanding of patient experiences</strong></td>
</tr>
<tr>
<td>Enhance understanding of treatment benefits and/or risks – acute and survivorship</td>
</tr>
<tr>
<td>For counselling during treatment decision-making</td>
</tr>
<tr>
<td>Characterize under-evaluated populations – e.g., rare conditions, socially disadvantaged, lacking health literacy, CALD</td>
</tr>
</tbody>
</table>

Au et al, Expert Review Pharmacoeconomic Outcomes Research 2010

---

**THE STUDY PROTOCOL**

**TIP: a comprehensive study protocol is the best way to start**

- outlines procedures for good conduct of the study
  - all relevant requirements so the study can be implemented uniformly by all sites and staff
  - detailed and clearly worded
  - success or failure of your study may depend on how well the protocol is designed and written
  - Chan et al. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials
- These general points apply just as much to PRO assessment as to any other aspect of a study
- Why and how PRO will be assessed
  - Mixture of science and logistics
- How well the PRO assessment methods are planned and described in the protocol will be a major determinant of the quality of PRO data, and any evidence/papers arising

**First Q: Is QoL/PRO primary or secondary endpoint?**

- If primary endpoint, you will need to provide much more detail than if secondary endpoint
- But you’ll be able to dedicate more space to it
- Need to think through all details of *study design* carefully
- If secondary endpoint, very little space available
- But still need to think through *study design* carefully
- QoL/PRO *study design* will have implications for budget (direct research costs, staff salary), timelines, and feasibility
In developing this checklist, we have drawn on:

- Our experience with ANZ Trials Group protocols
- Fairclough (2010): Design and Analysis of Quality of Life Studies in Clinical Trials
- SPIRIT-2013 Statement and Checklist

Provide a **Background and Rationale** for PROs in your protocol

- Rationale for measuring PRO in the study
  - What’s known / not known
  - Why it matters
- Provide an argument for the PRO variables considered relevant in your study
  - acute side effects
  - late effects – benefits & harms
  - symptoms & side-effects
  - psychological domains
  - core +/- other domains of functioning

Search, read and critically review the literature for your clinical context

E.g. Medline in Ovid
*<DISEASE>.ti
*<TREATMENT>.ab
*Quality of life.sh
*<KEY AUTHOR>.au

*Often one or more reviews already published
*Studies and reviews give you ideas about which PROs have been studied, why and how treatments affect them
*Which questionnaires have been used, response rates
Example PLUNG: Rationale from protocol

A number of randomised controlled trials (RCTs) comparing different palliative RT regimens in locally advanced NSCLC have been performed (9-20). Two systematic reviews of these data have been undertaken. A Cochrane review (21) of 14 RCTS involving 3576 evaluable patients found that palliative RT achieves reasonable rates of symptom control (haemoptysis, cough, pain, dyspnoea).

PLUNG - Rationale (cont.)

A second systematic review (22) confirmed the equivalence of specific symptom palliation but reported that, in comparison with lower dose schedules, higher dose schedules resulted in: a greater likelihood of symptom improvement on the total symptom score, a longer duration of symptom relief, an improvement in 1 year survival (26.5% vs. 21.7%, p =0.002) and a higher incidence of toxicity, predominantly oesophagitis.

Write clear Objectives and Hypotheses for PROs in your protocol

- The detail of the stated objective(s) should include:
  - P = Patient population
  - I = Intervention
  - C = Control or comparator
  - O = Outcomes, i.e. relevant PRO(s)
  - T = Time frame

Look for these in the following example...

PLUNG Objectives

Primary Objective: To compare, in this group of patients, high dose palliative radiotherapy (HDPRT) versus concurrent chemotherapy and HDPRT (C-HDPRT), with respect to:

- The relief of dyspnoea, cough, haemoptysis and chest pain as assessed by change in total symptom burden from baseline to six weeks after the completion of treatment.

- Response for each component symptom separately (dyspnoea, cough, haemoptysis, chest pain)
PLUNG Objectives

Secondary Objectives: To compare the two regimens in terms of:
- Dysphagia during treatment
- Thoracic symptom response rate
- Duration of thoracic symptom response
- QOL
- Progression-free survival
- Overall survival

WHICH QUESTIONNAIRE(S) TO USE?

Selecting a PRO Measure (PROM)

HELP!!! Too many to choose from!

Don’t panic!
The trick is to put the horse before the cart.

Decide on which PROs matter first, then choose the best questionnaires to measure them
Choosing the best questionnaire for your study

- First, review the specific PROs of interest, as stated in your objectives/hypotheses
- Look for questionnaire(s) that measure these and are known to be valid and reliable
- Review the content of candidate questionnaires
  - What issues are covered by the items (questions)?
  - How items are combined into multi-item scales?
- Choose the questionnaire that best matches your target PROs

Guidance on choosing QOL/PRO measures


Luckett & King. Choosing patient-reported outcome measures for cancer clinical research - practical principles and an algorithm to assist non-specialist researchers. EJC 2010.

Online Resource

Psychometric properties

- The measures you select must be psychometrically sound
  - As demonstrated, quantified in validation papers
- You must provide citations for validity in your protocol
- It's good to provide citations also for:
  - Reliability (internal consistency Cronbach’s alpha, test-retest)
  - Responsiveness to change (if available)
  - Interpretability / minimally important difference (MID)
**Patient burden**

How many questions (items) is too many?

- No longer than 20 mins (Basch E et al, JCO 2012)
- Calculate based on 12 sec per item
  - E.g. QLQ-C30, 30 items, 360 sec = 6 min
- Respondents may complete long batteries at baseline, but may be less inclined to complete repeatedly at follow-up assessments

---

**E.g. PLUNG – justification of measures**

Three QOL measures will be used: patient self completed QLQ-C30 (41) and QLQ-LC13 (42) and physician rated Spitzer Index (43).

The QLQ-LC13 was chosen as it contains the four most common and clinically important symptoms in this patient population (cough, dyspnoea, haemoptysis, chest pain), which constitute the primary endpoints in this study.

It also contains the main side-effect of radiotherapy (dysphagia).

---

**PLUNG – proximal v distal**

... It (the QLQ-LC13) is a module, designed to be used with the EORTC core quality of life questionnaire, QLQ-C30 (42). The latter will provide more distal measures of the impact of therapy on functioning (physical, role, emotional, social, cognitive), plus a global measure of QOL.

NOTE: together 43 items, ~ 9 mins to complete

---

**More PRO design considerations:**

who, how, where, when
PRO administration considerations

• Who?
  - Ideally all study participants, minimise exclusions
  - In practice, some limitations (language, literacy, physical, cognitive) may make self-assessment infeasible

• Where?
  - Clinic is a convenient place for research staff, but also consider if it is informative for all key timepoints
  - Home may be a more suitable place, e.g. survivorship studies

WHERE AND HOW TO ASSESS?

The PLACE of completion + HOW the subject completes the questionnaire

Mode of administration

Hard-copy questionnaires

**Pros**
- Convenient – patients complete while in clinic
- High completion rates
- Complete and collect questionnaires in one place → reduce risk of losing forms

**Cons**
- Administering adds work to already full site managers'/nurses' workload
- Data entry by hand is time-consuming and prone to human error
- Data entry by scanning has set-up costs
- When completed at home, returning by post is costly and risk of non-return or loss in mail

Online questionnaires

**Pros**
- Links to questionnaires and reminders can be sent by email
- Data entry and checking is automated
- Reduced costs, quicker return, efficient data management
- Optimise use of skip logic (IF Yes THEN more detail; IF No, THEN Skip)
- Allow 'Computer-adaptive test' (CAT) methods and scoring, e.g. PROMIS

**Cons**
- Patients can ignore email reminders
- Requires computer literacy & online access
- Technical errors (server or connectivity issues)

**Touchscreens & tablets**

**Pros**
- Feasible in oncology clinics
- Screening for distress in cancer
- Same benefits as for online and in-clinic data collection

**Cons**
- Does not eliminate risk of technical fault
- Expenses include staff, tablets/iPads, software
- Still need clinic staff to administer, although perhaps fewer administration tasks than hard-copy

**Computer-assisted telephone interview**

**Pros**
- Can assess at times not linked with clinic visits
- Enhances completion rates
- Allows inclusion of respondents with low English or computer literacy
- Good for slightly more complex cognitive tasks, e.g. preference-based measures
- Data entry done at time of data collection
- Good if research has been contracted out to CRO

**Cons**
- Costly: postage + phone + software + staff costs
- Time consuming

**MOA Systematic Review & Meta-analysis**

- Strong evidence that there is no bias:
  - paper vs electronic self-complete
  - self-complete vs assisted MOA
- Self-complete paper and electronic MOA can be used interchangeably in clinic and home settings.
- Self- and assisted-completion produce equivalent scores overall, although heterogeneity may be induced by setting.
- These results support the use of mixed MOAs within a study, which may be a useful strategy for reducing missing PRO data.


**When to assess?**
Timing of PRO assessments

- Getting the timing right matters as much as selecting the most appropriate instrument
- What to think about
  - acute and late effects
  - trajectories over time
  - timing and frequency of PRO assessment
  - acceptable time windows
  - assessment after discontinuation of protocol therapy in clinical trials, after recurrence in clinical trials and survivorship studies

Baseline

- Always! Why?
  - There is always a lot of between-person variation in PRO measures which persists over time
  - so adjustment for baseline is a good way to improve power of treatment comparisons
- What is a sensible baseline?
  - Clinically – pre-treatment, but what if a series of treatments?
  - Logically feasible?
- Intervention studies - before intervention starts
  - RCT: baseline PRO assessment as a criterion for randomisation
- Survivorship studies - at recruitment

Follow-up assessments

WHEN? HOW OFTEN? WHEN TO STOP?

- What is the expected trajectory of treatment effects?
  - Clinicians’ experience
  - Literature review
  - Pilot studies
- Intervention studies:
  - at end of treatment,
  - For medical treatment: how long to recovery from toxicity?
  - For psychosocial interventions - how long does benefit persist?
- Survivorship studies:
  - Typically periodic, e.g. every 3 months
  - Consider defining triggers for supplementary event-driven assessments, e.g. disease recurrence
- How often and when to stop post-intervention assessments
  - Feasibility – patient burden and staff salary
  - Implications for budget and timeline

Hypothetical: When would you measure?
Duration of PRO follow-up

- How long should PRO follow-up continue?
- If primary outcome is survival to 5 years, may be tempting to do annual PRO assessment to 5 years
- Problem – drop-out
  - Worse with advanced disease
- What is the expected attrition rate, given your target population?
- At what point is PRO data of diminished sample no longer of interest?
  - Expected median survival may provide a good guideline for the end of planned PRO assessment

Example: PLUNG

The primary endpoints are:
- the change in the Intrathoracic Symptom Burden Index from baseline to 6 weeks after end of treatment

The six week time-point has been chosen because this is when the maximal symptomatic benefit is expected.
Time windows

- Time period after the target event (e.g., surgery, 2nd week of radiotherapy, end of radiotherapy) within which effect of interest will be observed and not diluted
- Should be defined in the protocol

Example: PLUNG

The recommended acceptable time limits for completion of QOL assessments are:

<table>
<thead>
<tr>
<th>Assessment Time Point</th>
<th>Beginning of Acceptable Time Limit</th>
<th>End of Acceptable Time Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pre-Registration</td>
<td>Date of giving consent</td>
<td>Date patient made aware of treatment arm to which they have been assigned</td>
</tr>
<tr>
<td>2. Pre-Treatment</td>
<td>Day after patient aware of treatment arm assignment</td>
<td>Date of first radiation dose</td>
</tr>
<tr>
<td>3. During Treatment</td>
<td>Day of first treatment</td>
<td>Before day of next treatment</td>
</tr>
<tr>
<td>4. End of treatment</td>
<td>Day of end treatment</td>
<td>Five days after end treatment</td>
</tr>
<tr>
<td>5. 2 weeks after EOT</td>
<td>2 weeks after EOT</td>
<td>3 weeks after EOT</td>
</tr>
<tr>
<td>6. 6 weeks after EOT</td>
<td>6 weeks after EOT</td>
<td>Eight weeks after EOT</td>
</tr>
<tr>
<td>7. 8 months after EOT</td>
<td>8 months after EOT</td>
<td>Nine months after EOT</td>
</tr>
</tbody>
</table>

Individual cases

Actual vs planned return of Q’aires

Shaded area = time window
Are missing PRO data a problem?

Example 1: Adjuvant Breast Cancer Trial
- Experimental 16-week dose-intensive therapy vs. conventional 24-week therapy
- Hypothesis: Experimental therapy would result in superior disease-free and overall survival, but with more severe physical symptoms and inconvenience
- Timing of HRQoL assessments: 3 – before, during and after treatment

Example 1: Adjuvant Breast Cancer Trial
Rates and reason for missing data, n=200 recruited

<table>
<thead>
<tr>
<th>Status/Reason</th>
<th>Before Therapy</th>
<th>During Therapy</th>
<th>After Therapy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed</td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td></td>
<td>191 (96)</td>
<td>180 (90)</td>
<td>173 (87)</td>
</tr>
<tr>
<td>Too late for baseline</td>
<td>5 (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No longer receiving therapy</td>
<td>-</td>
<td>1 (1)</td>
<td>-</td>
</tr>
<tr>
<td>Patient refused</td>
<td>1 (1)</td>
<td>2 (1)</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Staff oversight</td>
<td>0 (0)</td>
<td>6 (3)</td>
<td>6 (3)</td>
</tr>
<tr>
<td>Patient too ill</td>
<td>0 (0)</td>
<td>11 (5)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Other, Early off therapy</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>11 (6)</td>
</tr>
<tr>
<td>Other, Not specified</td>
<td>3 (2)</td>
<td>0 (0)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Not documented</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>3 (2)</td>
</tr>
</tbody>
</table>


Example 2: Advanced NSCLC
Rates and reasons for missing data, n=525 recruited

<table>
<thead>
<tr>
<th>Status/Reason Missing</th>
<th>Baseline</th>
<th>6 weeks</th>
<th>12 weeks</th>
<th>6 months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>Completed</td>
<td>513 (98)</td>
<td>362 (75)</td>
<td>276 (63)</td>
<td>183 (52)</td>
</tr>
<tr>
<td>Refusal</td>
<td>0 (0)</td>
<td>34 (7)</td>
<td>46 (10)</td>
<td>16 (3)</td>
</tr>
<tr>
<td>Patient feels too ill</td>
<td>0 (0)</td>
<td>3 (1)</td>
<td>3 (1)</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Staff felt patient too ill</td>
<td>0 (0)</td>
<td>15 (3)</td>
<td>15 (3)</td>
<td>11 (3)</td>
</tr>
<tr>
<td>Staff oversight</td>
<td>7 (1)</td>
<td>35 (7)</td>
<td>59 (13)</td>
<td>56 (15)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (0)</td>
<td>13 (2)</td>
<td>13 (3)</td>
<td>20 (4)</td>
</tr>
<tr>
<td>Unknown</td>
<td>4 (1)</td>
<td>19 (4)</td>
<td>19 (4)</td>
<td>20 (6)</td>
</tr>
<tr>
<td>Patients expired</td>
<td>0 (0)</td>
<td>41 (8)</td>
<td>84 (16)</td>
<td>170 (32)</td>
</tr>
<tr>
<td>Total expected</td>
<td>525</td>
<td>484</td>
<td>441</td>
<td>355</td>
</tr>
</tbody>
</table>

Example 3: Advanced renal cell carcinoma trial

Completion rates, n=200 recruited

<table>
<thead>
<tr>
<th>Schedule (weeks)</th>
<th>0</th>
<th>2</th>
<th>8</th>
<th>17</th>
<th>34</th>
<th>52</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surviving patients</td>
<td>200</td>
<td>199</td>
<td>191</td>
<td>168</td>
<td>135</td>
<td>112</td>
</tr>
<tr>
<td>PROs completed</td>
<td>192</td>
<td>171</td>
<td>140</td>
<td>83</td>
<td>41</td>
<td>14</td>
</tr>
<tr>
<td>% of surviving patients</td>
<td>96</td>
<td>96</td>
<td>73</td>
<td>49</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>% of total patients</td>
<td>96</td>
<td>86</td>
<td>70</td>
<td>42</td>
<td>20</td>
<td>7</td>
</tr>
</tbody>
</table>


Missing PRO data: “informative missingness”

The impact of megestrol acetate (MA) on HRQoL was assessed in an RCT of patients with advanced endocrine-insensitive cancer; 240 patients were randomized double-blind to MA or placebo for 12 weeks. As early as 4 weeks after randomization, QOL assessments were available on only 160 (67%) of patients or 77% of patients still alive. Health status was significantly worse among patients not completing these assessments as judged by the clinician, based on clinician-rated QOL, toxicity or clinical deterioration.


Example 2: Ovarian cancer cohort study

FACT-G, higher score = better QOL

Graph is stratified by number of assessments completed

Those who drop out early start out worse and have steeper declines

Reporting missing data

NCIC QoL committee guidance, Osoba et al 2005

Reporting missing data

NCIC QoL committee guidance, Osoba et al 2005

Fig. 1. Proportions of health-related quality of life (HRQOL) questionnaires completed at baseline and during the study. The metis/hatched proportions show a marked diminution of questionnaires over time with only 7% completed by month 5. However, the missing/proportion retained high or >90% for all but month 4 of the study (adapted from Fantus, K. and colleagues [15].)
Are missing PRO data a problem?

**YES**

The missing PRO data problem

- Loss of power
- Bias
- Compromised external validity (generalizability) of the results
- Think about PRO missing data at **all stages of research**:
  - Protocol, activation, implementation/quality assurance, analysis and reporting

... (Continued on next page)
Strategies to prevent **avoidable** missing data

1) **STUDY DESIGN**

- Ensure protocol & site guidance is complete for PROs
- Timing, place and method of PRO assessments is well-thought out (Feasibility, practicality, convenience)
- Choose measures carefully – all that matters and no more
- Minimise patient burden - shorter, simpler questionnaire(s)
- Make baseline PRO completion a pre-randomisation eligibility condition
- Specify “time windows” for each assessment
  - For missed assessments: specify “makeup” procedures
- Plan to continue PRO assessment of patients who dropout of main trial

2) **IMPLEMENTATION – STAFF**

- Train staff: science & administration of PROs
- Start-up presentation – include slides on PRO, importance of good compliance and quality assurance basics
- Get clinicians and nurses on side – encourage PRO champions at each site
- Appoint a central PRO co-ordinator
  - Nominated person at each site, responsible for PRO assessments
- Site staff must approach all eligible on-trial pts - must be pt’s decision to complete/not to complete
- Ensure all participating sites have resources to comply - FUNDING

3) **IMPLEMENTATION – PATIENTS**

- Patient-related factors
  - Must be pt’s decision to complete/not to complete
  - Patient attitude – ensure patients understand the purpose of PRO assessments
  - Patient information sheets
  - Direct communication – recruiting staff/those administering PRO assessments
  - Offer to send results of study to patients when completed
Strategies to prevent *avoidable* missing data

2) IMPLEMENTATION: Quality assurance processes

- Record and monitor rates and reasons for missing data in a standardised and systematic way
- Give sites feedback on rates and reasons for missing data at their site
  - QOL Office Completion and Missing Data
  - Available in editable form on QOL Office website

The PRO CoMiDa Form

*Patient Reported Outcome Completion and Missing Data Form*

- A data management tool
- Aims to provide standardised documentation of the completion or reasons for non-completion of PRO assessments by patients in a clinical trial/study.
- Crucial for quality assurance - since missing data are the greatest threat to the integrity and interpretability of PRO data.
- Completed by the Data Manager/Research Nurse – or whomever is responsible for PRO data collection.
  - This person may be located at the site where PRO data is collected (e.g. at clinic), or centralised (e.g. when PRO assessments are completed at home).

**PRO CoMiDa Form (in parts)....**

- **Today’s date:** __/__/____
- **Site name:** __________________________
- **Site number:** ____________
- **Patient ID:** ________________________
- **Patient’s Initials:** __________________
- **Patient’s date of birth:** __/__/____

Select the current PRO Assessment time point:

- **Baseline**
- **Cycle 2**
- **Cycle 3**
- **Cycle 4**
- **Cycle 5**
- **End Treatment (EOT)**
- **3 month post-EOT**
- **1 year post-EOT**
- **3 year post-EOT**
- **1 month post-EOT**
**PRO CoMiDa Form (in parts)....**

1. Were the PRO forms completed at this scheduled assessment?

   Please complete each box below with one of the following codes:
   - 1 = Yes
   - 2 = NA (not required at this timepoint)
   - 3 = No

2. Did the patient require any assistance in completing the questionnaire?

   - No
   - Yes. Please describe: ________________________________

3. How were the questionnaires administered?

   - At clinic
   - By telephone
   - Online
   - Other: ______

4. What language were the questionnaires completed in?

   - English
   - Other: please specify: ________________________________

   If the patient completed ALL questionnaires required for this scheduled assessment, you have completed this form. If any questionnaires were MISSED (i.e. if you answered ‘3’ to any of the questionnaires in question 1 above), please continue.

5. Please select the most appropriate reason for non-completion of the questionnaire(s).

   - Patient received the questionnaire(s), but did not return them
   - Patient refused to complete questionnaire
   - Unable to contact patient
   - Patient missed appointment of scheduled assessment
   - Patient withdrew from study
   - Institution forgot to administer questionnaire
   - Institution administered incorrect questionnaire
   - Online questionnaire malfunction
   - Patient has passed away (tick ‘Yes’ for Q6)
   - Other. Please specify: ___________________________________________________________________

6. Is the reason for non-completion (as stated above) related to the patient’s illness?

   - Yes
   - No

   Notes: ___________________________________________________________________________________

I have reviewed the PRO CoMiDa Form and PRO Forms. All forms are complete or an explanation is given for any missing

Person completing this form: Name:______________ Signature: ___________________ Date: ____/____/_______

**PRO SAMPLE SIZE**
Sample size considerations (Walters 2009)

- **If PRO is the primary outcome:**
  - Standard methods for:
    - Continuous outcomes e.g. mean (MID) difference between trial arms
    - OR
    - Proportions e.g. % deteriorated by MID
  - Type 1 and II error rates as per usual
  - Minimally important difference (MID)

- **If PROs secondary outcome:**
  - Primary is typically a survival endpoint, say this will provide sufficient power for PROs

Sample size considerations (cont.)

- Anticipate attrition/drop-out — reduces power.
  - Need to inflate target sample size accordingly
  - Implications for budget and timeline
- Type 1 error rate, “false positives”, alpha, confidence level
  - Typically alpha=0.05, 95% confidence
  - If > 1 PRO is of key interest, share need to “share alpha”
  - Bonferroni (conservative), Hochberg (allows for correlations among PROs)
Questionnaire Scoring

Questionnaires usually have scoring guidelines
How to calculate sub-scales (domains)
  QoL: Physical, emotional, functional, social
How to handle missing items
  e.g., half mean imputation rule
How (if at all) to standardise
  e.g., instead of a possible range of 4-26, transform to 0-100
Reference these

If you have a good reason NOT to use standardised scoring ...
Describe:
  ➢ Which items
  ➢ Which timepoints
  ➢ Justifications
  ➢ Definitions

Caution: validity and reliability no longer assured
Include a validation analysis as part of the study’s statistical analysis plan

Example DIY PRO: PLUNG

The Intrathoracic Symptom Burden Index is defined, for the purposes of this study, as the average of 4 components based on items from EORTC QLQ LC13: dyspnoea (3 items, Q33-Q35), cough (Q31), haemoptysis (Q32) and chest pain (Q40). The scores for the three dyspnoea items will be averaged. The average of this figure and the scores for cough, haemoptysis and chest pain will be standardised to a 0 to 100 scale to give the Intrathoracic Symptom Burden Index (ISBI).

\[
\text{ISBI} = 100 \times \frac{[(Q_{33} + Q_{34} + Q_{35})/3 + Q_{31} + Q_{32} + Q_{40}]/4 - 1]/3
\]

where, \(Q_{33}\) is the score (1, 2, 3 or 4) for question Q33, etc.

Analysis Metric

What are you going to do analyses on?
  ➢ PRO scale scores at each timepoint
  ➢ Change from baseline
  ➢ Responder (yes/no)
  ➢ Area under the curve
Endpoints: PLUNG

The primary endpoints are:
• the change in the Intrathoracic Symptom Burden Index from baseline to 6 weeks from the end of treatment; and
• the changes in each of the four symptoms, dyspnoea, cough, haemoptysis and chest pain from baseline to 6 weeks from the end of treatment.
• The six week time-point has been chosen for the primary endpoints because this is when the maximal symptomatic benefit is expected.

Derived variables: PLUNG

• Area under curve (AUC) of dysphagia (LC13, Q37) during treatment, derived from changes from baseline at weeks 3, 2 and 1 prior to the end of treatment and at the end of treatment, using the trapezoidal method.

Responder definitions

Translate the PRO into a binary outcome defined by whether the participant achieved a clinically important response e.g., improvement of at least the minimally important difference in QoL. Participants who dropout or die are coded as a treatment failure, and missing data is thus reduced or eliminated.

Responder definition: PLUNG

Thoracic symptom response and response duration. Thoracic symptom response will be defined as a reduction of the Intrathoracic Symptom Burden Index by at least 10 units (on the standardised 0-100 scale) from baseline.
Responder definitions – potential drawbacks

- loss of power
- cut-off values to define a response are somewhat arbitrary
- misclassification bias of individuals
- biased estimation (how big? what direction?), and the potential to mislead
A response may be due to regression to the mean, measurement error, the natural history of disease, or other concurrent therapies.

Minimal Important Difference

The MID is important for
- Sample size calculation
- Responder definitions
- Interpretation


Suggested readings re MIDs

- King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Review of Pharmacoeconomics and Health Outcomes 2011.

Statistical Analysis

The statistical analysis should align with your objectives and measures

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Primary</th>
<th>Secondary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures</td>
<td>Primary</td>
<td>Secondary</td>
</tr>
<tr>
<td>Analysis</td>
<td>Primary</td>
<td>Secondary</td>
</tr>
</tbody>
</table>

Sample size calculation is a function of all three!
Statistical Analysis

Include statistical methods for
- primary outcomes
- secondary outcomes
- sensitivity analyses, particularly as regards to missing data

A more detailed plan, the Statistical Analysis Plan (SAP) may be developed after the protocol (while the data are being collected and before they are ready to analyse)

Advantage of separate SAP is that changes to it do not require protocol amendment through ethics committees

Statistical Analysis

Care should be taken when defining which subjects should be excluded from analysis

Analyses are often limited to subjects with at least two QOL assessments (baseline + one follow-up)
- Excludes patients who drop out of the trial early
- If the proportion excluded is small, it won’t matter
- If substantial, this could have a substantial impact on results and generalisability
  - Comparison of those excluded and those included
  - Consider analysis options

PLUNG

Analysis of the primary objective

The primary objective will be assessed initially by comparing the mean change in Intrathoracic Symptom Burden Index from baseline to 6 weeks from end of treatment between arms, adjusting for baseline index. The analysis will be performed using a linear regression model.

PLUNG

All patients with a valid baseline QOL assessment and at least one follow-up QOL questionnaire will be included in the analysis. The baseline questionnaire will be considered valid if filled out and dated by the patient in the two weeks prior to the starting date of trial treatment.
Assessment of missing data

Rate and reasons for missing data should be assessed
- IF RCT - by trial arm
- at each timepoint

PLUNG: Reasons for missing baseline and follow-up QOL questionnaires will be assessed. A table of QoL compliance rates (QoL assessments completed as a percentage of expected) by arm will be reported.

Handling Missing Data

- Not accounting for those who have dropped out is likely to
  - overestimate the HRQoL and function in your trial / study / patient population
  - underestimate the symptoms and toxicity
- How will missing data be handled analytically?
- What assumptions are being made about the type of missing data?
- Will the analytical approach give unbiased estimation?
- What types of sensitivity analyses are planned?

Little 2012

Types of missingness: Stat_speak

- Missing completely at random (MCAR)
  - Missingness is not related to past or present (but unobserved) outcomes.

- Missing at random (MAR)
  - Missingness IS related to past outcomes but not present (but unobserved) outcomes (as long as past is taken account of).

- Missing not at random (MNAR)
  - Missingness is related to present (but unobserved) outcomes, even after taking account of past.

NCIC QoL committee guidance (Osoba 2005)

Fig. 3. Proportion of hand-held quality of life (HRQoL) questionnaires completed at baseline and during the trial. The proportion of baseline response shows a marked reduction of questionnaires over time with only 25% completion by month 3. However, the proportion of completed questionnaires reassured high at 70% for all but month 4 of the trial. (adapted from: Osoba and colleagues, 2005).
Implications of missing data

- MCAR => loss of power only
- MAR => loss of power and bias for some methods
- MNAR => loss of power and bias for most methods

Handling missing data: imputation

- Simple imputation: one value is filled in
  - Mean, last observation carried forward (LOCF), linear interpolation/extrapolation
  - These methods are not valid because
    - They don’t account for uncertainty in the data
    - Standard errors are too small which means confidence intervals are too small
    - They make unrealistic assumptions!

Handling missing data: alternatives to simple imputation – MLE and MAR models

- Maximum likelihood estimation (MLE)
  - Estimation method used in mixed model repeated measures (MMRM), aka hierarchical linear models, random effect models, multi-level models, generalised estimating equations (GEE)
  - Assume missing PRO data are missing at random (MAR)
    - Unbiased if data are MAR and the model has been specified correctly (Carpenter & Kenward 2008)
    - Recommended for use in the regulatory environment because their generality protects against misspecification to all but the missing data mechanism (Mallinkrodt 2003)

Try to avoid missing data in the first place!
Missing data: MNAR models

- If there is a substantial amount of missing data (>10%? no hard and fast rules), and one suspects that data may be MNAR — pattern mixture models, selection models, shared parameter models, and joint multivariate models
- Very complex! Do not attempt without a certified statistician

Carpenter & Kenward 2008
Bell & Fairclough 2014
Fairclough 2010

Handling missing data: Multiple Imputation

- Suppose there are two QoL observations on each participant, Y1 (which is completely observed) and Y2 (which has some data missing).
- We can learn about the association between Y1 and Y2 from those participants who have complete data, via the imputation model, and then use this to fill in the missing values of Y2.
- The assumption that is being made here is the relationship between Y1 and Y2 is the same for those who complete and those who do not.

This assumption can not be tested.

Multiplicity / Multiple testing

- Multiple testing can cause an increased Type I error rate
  - “false positives”
  - finding a significant result by chance
- Multiplicity arises because:
  - there are often multiple outcomes
    - e.g. for HRQOL: overall QoL, physical, social, emotional, and functional well-being, plus symptoms
  - PROs are often measured at more than one time-point

Multiplicity

There are three main strategies:

- Limiting the number of hypothesis tests by focussing on key domains and time points
- Using summary statistics, such as area under the time-curve
- Multiple comparison procedures, which includes adjustment to the alpha level of the tests...
Type I error rate control

**Bonferroni:**
If there are $k$ tests, the criterion of significance for each is $0.05/k$.
e.g., for 20 tests, criterion of significance = $0.05/20 = 0.0025$

**Hochberg** (1988), Benjamini & Hochberg 1995:
e.g., if there are 15 tests:
the largest $p$ value is compared to $0.05$
the second largest to $0.05/2$
...the 12th largest to $0.05/12$, etc.
In this sequence, as soon as one is significant, the rest will be.

From the statistical analysis sections:

**PLUNG:** The Hochberg method will be used to assess the four endpoints while maintaining the overall type I error at 5% or less.

**HPV:** Adjustment for multiple comparisons of PRO differences at specific times will be undertaken using the Hochberg method. No formal adjustment will be used for other secondary or exploratory analyses. In general, two-sided $P$-values will be used and 95% confidence limits for all important endpoints will be reported.

**PRO Analysis – challenges and solutions**

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Solutions</th>
</tr>
</thead>
</table>
| Numerous PRO variables and time-points | • Pre-specify key PROs and time-points to “spend alpha on”, treat others as “exploratory”
• Adjust $p$-values – Bonferroni, Hochberg |
| Informative missing data | • Evaluate impact of missing data (rates and reasons)
• “Responder analysis” – devise criteria for responders versus non-responders, e.g. a composite endpoint that equates observed deterioration in PROs with unobserved but likely deterioration (still alive but too unwell to complete) and death
• MLE and multiple imputation using auxiliary data |
| Repeated measures over time, correlated | • Condense the observations for each patient into a summary measure (e.g. Mean, AUC, etc.)
• Models that accommodate correlated data
• Adjust for baseline PRO levels |

---

**Example – Stockler et al, JCO 2014**

*Journal of Clinical Oncology*  
**Original Report**

Patient-Reported Outcome Results From the Open-Label Phase III AURELIA Trial Evaluating Bevacizumab-Containing Therapy for Platinum-Resistant Ovarian Cancer

Analyzing patient-reported outcome (PRO) endpoints in advanced cancer clinical trials: responder analysis versus mixed model for repeated measures (MMRM) illustrated with results from the AURELIA ovarian cancer trial.
Primary endpoint: Progression-free survival
Secondary endpoints: patient-reported abdominal symptoms, HRQoL

AURELIA primary endpoint result: Progression-free survival
Median duration of follow-up: 13.3 months (CT arm) vs 13.0 months (BEV + CT arm)

<table>
<thead>
<tr>
<th>Time (months)</th>
<th>CT (n=182)</th>
<th>BEV + CT (n=179)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>166 (91%)</td>
<td>126 (72%)</td>
</tr>
<tr>
<td>3.4</td>
<td>(2.2–3.7)</td>
<td>(5.7–7.9)</td>
</tr>
<tr>
<td>HR (unadjusted)</td>
<td>0.46</td>
<td>(0.38–0.60)</td>
</tr>
<tr>
<td>(2-sided, unadjusted)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

Estimated probability

AURELIA - PRO hypotheses
- Primary PRO hypothesis focused on abdominal/GI symptoms, which are considered the predominant symptoms of recurrent ovarian cancer
- A higher proportion of women allocated to BEV–CT would experience a substantial (≥15%) improvement in abdominal/GI symptoms at week 8/9, QLQ-OV28 Abdo/GI scale

Primary PRO analysis – Responder analysis
- For the primary PRO analysis:
  - The abdominal/GI symptoms endpoint was based on items 31–36 of QLQ-OV28 (abdominal pain, feeling bloated, clothes too tight, changed bowel habit, flatulence, fullness when eating)
  - excluding the non-specific item 37 (heartburn/indigestion)
  - Responder analysis – composite endpoint
    - patients who experienced <15% improvement on QLQ-OV28 abdo/GI symptom subscale were coded as a treatment failure.
    - patients who missed questionnaires at week 8/9 due to PD or death, or who had stopped study participation, were also coded as a treatment failure.
  - WHY? Clinically sensible way to deal with missing data, because unlikely that those patients had experience improvement in abdo/GI symptoms.
Sensitivity analyses I: thresholds and scoring

- Responder sensitivity analyses:
  - 10% threshold (vs 15%)
  - Only patients who had both baseline and week 8/9 PRO assessments (paired values or “complete case”) were included
- Sensitivity to non-standard scoring:
  - Abdo/GI scale was scored as per Greimel et al 2003 (i.e. 7 items, incl heartburn)

Sensitivity analyses II: MMRM

- Linear mixed model repeated measures (MMRM) analysis as an alternative approach to missing data
- Used to compare trial across all time points until PD/death
- Uses all available data
- Assumes missing data are “missing at random” (MAR) – unlikely?
- PROs analysed as continuous variable, adjusting for score at baseline, time & treatment-by-time interaction
- Time was considered as a categorical variable, patient as random variable, and an unstructured covariance structure was assumed.

AURELIA primary endpoint result: Progression-free survival

Median duration of follow-up: 13.9 months (CT arm) vs 13.0 months (BEV + CT arm)

<table>
<thead>
<tr>
<th></th>
<th>Median PFS, months (95% CI)</th>
<th>166 (91%)</th>
<th>135 (75%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT (n=182)</td>
<td>3.4</td>
<td>2.2–3.7</td>
<td>3.7–7.9</td>
</tr>
<tr>
<td>BEV + CT (n=179)</td>
<td>6.7</td>
<td>5.7–7.9</td>
<td>5.7–7.9</td>
</tr>
</tbody>
</table>

HR (unadjusted) (95% CI) Log-rank p-value (2-sided, unadjusted)

- 0.48 (0.38–0.60) <0.001

Compliance with PRO questionnaire completion

<table>
<thead>
<tr>
<th></th>
<th>CT (n=182)</th>
<th>BEV–CT (n=179)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients (% of evaluable patients)</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Baseline</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Week 8/9</td>
<td>92</td>
<td>91</td>
</tr>
<tr>
<td>Week 16/18</td>
<td>88</td>
<td>85</td>
</tr>
</tbody>
</table>

Denominator:
- CT 182 164 91
- BEV–CT 179 176 143
*Denominator (patients known to be progression-free) excludes patients whose disease progressed or who died or were lost to follow-up at least 14 days before the scheduled assessment date
Patients with ≥15% improvement (%)

<table>
<thead>
<tr>
<th></th>
<th>CT (N=182)</th>
<th>BEV-CT (N=179)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>12.7% [4.4–20.9]</td>
<td>12.0% [4.5–19.5]</td>
</tr>
<tr>
<td>10% cut-off</td>
<td>29.9% [11.5–20.6]</td>
<td>13.3% [4.5–22.1]</td>
</tr>
<tr>
<td>Complete cases</td>
<td>23.4% [11.5–20.6]</td>
<td>10.1% [4.5–22.1]</td>
</tr>
</tbody>
</table>

Numbers in square brackets represent 95% confidence intervals with the Hauck-Anderson continuity correction for the difference between arms.

Sensitivity analyses

AURELIA RESULTS: Primary PRO hypothesis (abdo/GI symptoms): Main and sensitivity analyses, week 8/9

AURELIA RESULTS: Abdominal/GI symptoms: Mixed-model repeated-measures analysis, weeks 8/9 to 30

Least-squares means estimate

AURELIA Secondary PRO endpoints QLQ-C30:
Patients with improvement from baseline at week 8/9

Patients achieving a ≥15% improvement from baseline, n/N (%)

<table>
<thead>
<tr>
<th>Subscale</th>
<th>CT (N=182)</th>
<th>BEV-CT (N=179)</th>
<th>Difference, % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical functional</td>
<td>3/170 (1.8)</td>
<td>20/167 (12.0)</td>
<td></td>
</tr>
<tr>
<td>Role functional</td>
<td>17/170 (10.0)</td>
<td>37/167 (22.2)</td>
<td></td>
</tr>
<tr>
<td>Emotional functional</td>
<td>20/168 (11.5)</td>
<td>39/164 (23.8)</td>
<td></td>
</tr>
<tr>
<td>Social functional</td>
<td>21/167 (12.6)</td>
<td>37/163 (22.7)</td>
<td></td>
</tr>
<tr>
<td>Global health status/QoL score</td>
<td>23/169 (13.8)</td>
<td>40/164 (24.4)</td>
<td></td>
</tr>
</tbody>
</table>

- Supported by cut-point & complete case sensitivity analyses
- Not supported by the MMRM analysis
AURELIA Conclusions

- For primary PRO endpoint, corroboration across a comprehensive & sensible sensitivity analysis suggest the results are robust
- Secondary PRO endpoints, MMRM did not support responder analysis, but data unlikely to be MAR – to what extent were MMRM estimates biased?
- The benefits of bevacizumab in AURELIA extended beyond the prolongation of PFS to include significant improvements in abdominal/GI symptoms and improvements in various aspects of functioning and global health-related QoL
- Bevacizumab combined with chemotherapy should be considered a new standard option in platinum-resistant ovarian cancer

INTERPRETING HRQOL & PROs

Key points

- Challenges to interpreting PROs and PROMs
- Some solutions
- Statistical significance versus clinical importance
- Minimally important difference (MID)

How we assign numbers to PRO scales

- “measurement” = “assigning numbers by rules” Stevens (Science 1946)
- Questionnaires have simple rules for assigning numbers
- Standard set of questions
  - a representative sample of all relevant items
  - operationalises the definition of HRQOL domain
- Standard set of response scales
  - assign numbers to perceptions
- Standard scoring algorithm
  - defines the measurement scale
  - each dimension score is the (weighted) sum of item scores
  - sometimes linear transform: eg 0-100 scale range
  - sometimes norm-based: (mean, SD) eg (50, 10)
SF-36 physical functioning scale

10 items
1. Vigorous activities – running
2. Moderate activities – vacuuming
3. Lifting or carrying groceries
4. Climbing several flights of stairs
5. Climbing one flight of stairs
6. Bending, kneeling, stooping
7. Walking more than a kilometer
8. Walking several blocks
9. Walking one block
10. Bathing or dressing

Response options:
1. Yes, limited a lot
2. Yes, limited a little
3. No, not limited at all

Transform raw scores into scale scores

SF-36 Manual & Interpretation Guide Table 9.1

Transformed scale =
\[
\frac{\text{actual raw score} - \text{lowest possible raw score}}{\text{possible raw score range}} \times 100
\]

Example: A physical functioning raw score of 21 would be transformed as follows:
\[
\frac{21 - 10}{20} \times 100 = 55
\]

Physical Functioning “Ruler”

Some observations about these scales?

- Ordinal, not interval
- Arbitrary – no fixed zero point
- For multi-item scales
  - Number of items determines number of distinct values on the scale
  - Meaning linked to specific values is difficult to determine
  - How to interpret change on a scale?
  - Ditto – at high versus low end of the scale?
  - How to interpret change across scales?
What’s the problem?
Breast cancer to illustrate: Impact on HRQoL

Lots of domains, lots of items, lots of scales, lots of results to digest and interpret

- The scales look similar
  - 0 to 100 range
  - Some higher = better, others higher = worse
- But each contains a different set of items
  - So not directly comparable
  - I.e. can’t necessarily say a mean diff of 10 for arm symptoms is the same as a mean diff of 10 for fatigue

Example: CMF v UFT in node-neg Br Ca
Results: EORTC QLQ-C30/BR23

Side-effects: higher score = more symptoms = worse QOL

Social functioning: higher score = better QOL

Watamabe et al. Oral uracil and tegafur compared with CMF as postoperative chemo in node-negative, high-risk breast cancer. JCO 2009.

Things to think about when interpreting these graphs

- Do all domains go in same direction?
- Statistical significance versus clinical importance
- Are observed effects big? Or small?
- Can we interpret “big” and “small” in the same way across different domains?
- What is the smallest difference that matters?
  — Minimally Important Difference (MID)

QOL results are often not interpreted in a way that is clinically meaningful

- In a recent review of the interpretation strategies for QOL from RCTs that use the cancer-specific EORTC QLQ-C30:
  — Clinical significance was addressed in only 38% (n=82 papers).
  — Where clinical significance was not addressed, reliance was usually on statistical significance.


Clinical vs Statistical Significance
The problem with statistical significance
(IF you don’t know a priori what the clinically important difference is)

95% CI around a mean difference of 10 points between two groups, SD=20, ES = 0.5, 80% power

QOL scale (0-100)

Sample size per group

p>0.05 ... As n increases, CI decrease, p decreases ... p<0.05 when crosses zero line

Even a very small difference can be statistically significant with a large enough sample size

How can we know whether an observed difference is clinically significant?

One approach is “Effect size”
Signal to noise ratio
Signal = mean difference
Noise = SD

ES = 79-70
18
0.5

Interpreting Effect Sizes

• ES is a very appealing metric, because its interpretation is the same for all scales
  – Mean difference (PROM-i units) is interpreted in terms of variation among individuals in the sample (standard deviation, also PROM-i units)

• Cohen’s guidelines
  Statistical Power Analysis for the Behavioural Sciences 1988
  • Small ES = 0.2 one fifth of a standard deviation
  • Medium ES = 0.5 half of a standard deviation
  • Large ES = 0.8 four fifths of a standard deviation

• Very widely used
  – 0.5 is widely considered the upper limit of minimally important difference (MID)

What about the ‘minimally important difference’ or MID?
What is a MID?

- The terminology is confusing, several terms differing only slightly in definition
  - minimal clinically important difference (MCID)
  - clinically important difference (CID)
  - minimally important difference (MID)
  - minimally detectable difference (MDD)
  - subjectively significant difference (SSD)
- Summarised in King

Nuances of definition are of little consequence in the way these quantities are estimated and used.

How is MID determined?

- Anchor-based
  - Requires an "anchor" additional to the HRQL/PRO data
  - patient rating of change
  - clinical anchors
- Distribution-based
  - Calculated from the HRQL/PRO data
  - effect size = mean diff / SD
  - standard error of measurement SEM = SD * square root of 1-reliability

MID recommendations

- There is no universal MID, despite the appeal of the notion
- For a particular PRO instrument or scale, the MID is not an immutable characteristic, but may vary by population and context.
- At both group and individual level, the MID may depend on the clinical context and decision at hand, the baseline from which the patient starts and whether improving or deteriorating.
- Specific estimates of MIDs should therefore not be over-interpreted.
- For a given PRO scale, all available MID estimates (and their confidence intervals) should be considered, amalgamated into general guidelines, and applied judiciously to any particular clinical or research context.

Revisit example: CMF v UFT in node-neg Br Ca

Results: EORTC QLQ-C30/BR23

- Side-effects: higher score = more symptoms = worse QOL
- Social functioning: higher score = better QOL

Watanabe et al. Oral uracil and tegafur compared with CMF as postoperative chemo in node-negative, high-risk breast cancer. JCO 2009.
**CONSORT-PRO extension**

- Adapts and builds on the CONSORT statement – for PROs
- Some criteria only applicable to primary PRO endpoints
- Designed to help researchers publish research transparently
  - Often ADAPTED for use in studies EVALUATING reporting quality retrospectively
  - Findings: a lot of poor quality PRO reporting!
- Parallels QOL Office PROtocol Checklist items

**CONSORT-PRO**

1. The PRO should be identified in the abstract as a primary or secondary outcome
2. Scientific background and explanation of rationale - Including PRO assessment
3. The PRO hypothesis should be stated and relevant domains identified, if applicable
4. Participant eligibility criteria - required if PROs were used in eligibility or stratification criteria
5. Evidence of PRO instrument validity/reliability should be provided/cited, including the person completing the PRO and methods of data collection (paper, telephone, electronic, other).
6. How sample size was determined - Not required for PRO unless it is a primary study outcome
7. Statistical approaches for dealing with missing data are explicitly stated

---

**The CONSORT-PRO Extension**

Calvert 2013

Reporting of Patient-Reported Outcomes in Randomized Trials

The CONSORT PRO Extension

Melanie Calvert, PhD
Jean Blackley, MD
Douglas C. Hines, BSc
Dana A. Revicki, PhD
David Becker, PhD
Michael D. Brazicone, MD
for the CONSORT PRO Group

The CONSORT (Consolidated Standards of Reporting Trials) statement aims to improve the reporting of randomized controlled trials (RCTs); however, it lacks guidance on the reporting of patient-reported outcomes (PROs), which are often inadequately reported in trials, thus limiting the value of these data. In this article, we describe the development of the CONSORT PRO extension based on the methodological framework for guideline development proposed by the Enhancing the Quality and Transparency of Health Research (EQUATOR) Network. Five CONSORT PRO checklist items are recom
CONSORT-PRO

8. Available PRO data baseline and at subsequent time points should be made transparent
9. A table showing baseline demographic and clinical characteristics for each group, including baseline PRO data when collected
10. For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups. Required for PRO results
11. Results for each group, the estimated effect size, and its precision such as 95% confidence interval, for multidimensional PRO results from each domain and time point
12. Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory - Including PRO analyses, where relevant
13. PRO-specific limitations and implications for generalizability and clinical practice
14. PRO data should be interpreted in relation to clinical outcomes including survival data, where relevant

Cancer Australia / Univ Sydney
QOL Office Online Resources

PROtocoll Checklist
1 or 2-day workshops, please contact us if you are interested

CoMiDa form
CRF to record PRO form completion rates & reasons for missing data

Statistical analysis position paper

W: Pocog.org.au/qoloffice
E: qol.office@sydney.edu.au