Core Outcome Set for Behavioural Weight Management Interventions for Adults with Overweight and Obesity: STAndardised Reporting of Lifestyle Weight Management InTerventions to Aid Evaluation (STAR-LITE)

Ruth M. Mackenzie[1], *Louisa J. Ells*[2], *Sharon Anne Simpson*[3], *Jennifer Logue*[1]

[1] Institute of Cardiovascular and Medical Sciences, University of Glasgow

[2] School of Health & Social Care, Teesside University

[3] Institute of Health and Wellbeing, University of Glasgow

Keywords: core outcome set, adult behavioural weight management interventions, standardised reporting

Running title: Core Outcome Set for Adult Behavioural Weight Management Interventions (BWMIs)

Acknowledgements: The authors wish to thank all expert group members, as listed in Supporting Information 15, for their time and valuable input. We also thank Rob Froud and Jonathan Foss of Clinvivo for their Delphi expertise.

Corresponding author:

Jennifer Logue

Institute of Cardiovascular and Medical Sciences

University of Glasgow

126 University Place

Glasgow, G12 8TA

Jennifer.Logue@glasgow.ac.uk

Conflicts of interest: JL leads a joint working project between University of Glasgow, National Health Service (NHS) Greater Glasgow and Clyde, MSD and Astra Zeneca. The project also involved an educational grant from Janssen. JL received funding to attend a conference from Novo Nordisk. LJE has a part time secondment with Public Health England. LJE has also received funding from North Yorkshire County Council, PHE, Sport England, National Prevention Research Initiative (NPRI) and Chief Scientist Office (CSO) Scotland in the last 3 years.

Abbreviations: STAndardised Reporting of Lifestyle Weight Management InTerventions to Aid Evaluation, STAR-LITE; behavioural weight management intervention, BWMI; National Health Service, NHS; National Prevention Research Initiative, NPRI; Chief Scientist Office, CSO; National Institute for Health and Care Excellence, NICE; Scottish Intercollegiate Guidelines Network, SIGN; United Kingdom, UK; Public Health England, PHE; standard evaluation framework, SEF; body mass index, BMI; Core Outcome Measures in Effectiveness Trials, COMET; core outcome set, COS; Core Outcome Set-STAndards for Reporting, COS-STAR; key performance indicators, KPI; Research ANd Development, RAND; University of California Los Angeles, UCLA; inter-percentile range, IPR; inter-percentile range adjusted for symmetry, IPRAS; quality of life, QoL; Edmonton Obesity Scale Score, EOSS; haemoglobin A1c, HbA1c; United States of America (USA); Medical Research Council, MRC.

ABSTRACT

Background: Behavioural weight management interventions in research studies and clinical practice differ in length, advice, frequency of meetings, staff and cost. Few real-world programmes have published patient outcomes and those that have used different ways of reporting information, making it impossible to compare interventions and develop the evidence base. To address this issue, we have developed a core outcome set for behavioural weight management intervention programmes for adults with overweight and obesity.

Methods: Outcomes were identified via systematic review of the literature. A representative expert group was formed comprising people with experience of adult weight management services. An online Delphi process was employed to reach consensus as to which outcomes should be measured and reported, and which definitions/instruments should be utilised.

Results: The expert group identified 8 core outcomes and 12 core processes for reporting by weight management services. 11 outcomes and 5 processes were identified as optional. The most appropriate definitions/instruments for measuring each outcome/process were also agreed.

Conclusions: Our core outcome set will ensure consistency of reporting. This will allow behavioural weight management interventions to be compared, revealing which interventions work best for which members of the population and helping inform development of adult behavioural weight management interventions.

INTRODUCTION

Behavioural weight management interventions (BWMIs), known in the United Kingdom (UK) as tier 2 services, are the first line treatment for overweight and obesity[1-4]. International guidelines, including those of The National Institute for Health and Care Excellence (NICE)[1], Scottish Intercollegiate Guidelines Network (SIGN)[2], and the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society[3], outline the intervention components to be included in a behavioural weight management programme for adults. These components which include, calorie restriction, increased physical activity and behavioural change support, have proven efficacy in randomised controlled trials[3]. However, their implementation in practice is inconsistent. Indeed, mapping exercises in Scotland[4] and England[5] revealed wide variation in adult weight management services with regard to inclusion criteria, referral routes, delivery format, programme length and cost, despite the single-payer healthcare system. Furthermore, few adult BWMIs have published outcome data and where these data are published, results are often poor with low levels of programme completion and 'success', with a lack of longer term outcomes[6, 7].

When developing the guidance, 'Weight management: lifestyle services for overweight or obese adults'[1] in 2014, NICE identified a number of evidence gaps. These included, reliance on studies with short follow-up, collection of data at limited time points, small sample sizes, demographic samples that limit the ability to generalise, non-reporting of reasons for people dropping out and lack of evidence regarding the effect of population characteristics, such as age, gender and socio-economic status, on the effectiveness of a service. NICE specifically mentioned "variable outcome definitions" used in the clinical trials, which formed the supporting systematic review and meta-analysis, as a major barrier to developing evidence based guidance. As a result, they were left with many evidence gaps including "a lack of trials directly comparing lifestyle weight management programmes in the UK" and "a general lack of evidence on which specific components of a lifestyle weight management programme ensure effectiveness". This lack of an evidence base from both clinical trials and real-world services means that it is not possible to issue clear guidance as to which services are cost effective for which population groups.

Public health bodies in the UK have made efforts to try and address this issue; Public Health England (PHE)[8] created a standard evaluation framework (SEF) for weight management programmes[9]. However, PHE were unable to analyse data from real world interventions due to the heterogeneity of reporting, suggesting further guidance is required. This heterogeneity can be exemplified by reporting of weight loss which included, number of kilograms lost, percentage weight loss, average number of completers achieving 5% weight loss and body mass index (BMI)[5]. With regard to clinical trials, evidence suggests similarly heterogeneous reporting of outcomes[7].

It is acknowledged that the provision of treatments for obesity is severely limited across the world[10-14] and large gaps in the evidence of effectiveness may be contributing to this. An improved evidence base would allow intervention programmes to be commissioned and funded by health systems with the confidence of effectiveness. There is an urgent need to gain consensus on standardised outcome reporting to allow better comparison and meta-analysis of interventions to be performed across both real world and trial interventions. Therefore, the specific aim of this study was to use Delphi methodology to gain expert consensus opinion on the core outcomes that should be reported from BWMIs in real world clinical practice as well as within research studies, and on the outcome definitions/outcome measurement instruments that should be used in their evaluation. Core outcome set (COS) development has an established methodology[15] and COS represent the minimum that should be reported in all clinical trials of a specific condition, while also being suitable for observation research and audit; their use in clinical trials is supported by the UK National Institute of Health Research (NIHR)[16] as it allows trial results to be easily compared and combined. However, the development of a COS does not imply that research outcomes should be restricted to only those included in the COS. The development of these core outcome and definition/instrument sets for

BWMIs will ensure more consistency in the measurement of the effectiveness of weight management services, leading to a better evidence base from which to identify which services are effective across a range of settings.

METHODS

Ethics

Ethical approval for this study was received from the University of Glasgow College of Medical, Veterinary and Life Sciences Ethics Committee.

The project has been registered with the COMET (Core Outcome Measures in Effectiveness Trials) Initiative (http://www.comet-initiative.org/studies/details/1056) and a detailed methodology has been reported previously[17]. In reporting the development of our COS, we have adhered to the COS-STAR (Core Outcome Set-STAndards for Reporting) Statement (Table S1)[18].

Identification of outcomes

In order to develop a COS, a comprehensive list of outcomes for reporting from behavioural weight management interventions was generated. These outcomes were identified following review of studies included in the systematic review, 'The clinical effectiveness of long-term weight management schemes for adults' by Hartmann-Boyce *et al.* (2013)[7], conducted during the development of NICE guidance[1]. This review was updated to cover the time period 1/11/2012 until 30/09/17 using the same inclusion criteria (inclusion criteria and additional studies are outlined in Supporting Information 1). Both primary and secondary outcomes from studies were identified by two independent researchers and entered into a spreadsheet. Additionally, the PHE SEF[9], minimum dataset[19] and key performance indicators (KPI) document[20] were reviewed, again by two independent researchers, and any supplementary outcomes added to the aforementioned spreadsheet. Of note, the PHE SEF[9] was developed following focus group work with a wide range of stakeholders, including weight management staff, primary care staff, academics, commissioners and policy makers, and has been refined over 2 versions from 2009 to 2018.

Identification of outcome measurement instruments/outcome definitions

Analyses of studies identified during the systematic review by Hartmann-Boyce *et al.* (2013)[7] and our updated search (Supporting Information 1) allowed instruments and definitions for selected outcomes to be added to the data extraction spreadsheet by two independent researchers. This list was then examined by all study investigators and further suitable instruments/definitions added.

Participants

The core outcome and instrument set was developed by means of consensus from an expert group, recruited as outlined previously[17] and selected based on our sampling framework (Supporting Information 2) to ensure a representative sample and a pragmatic and patient-centred core outcome set. All experts recruited were from the UK.

For the stage 1 (outcome selection) Delphi process, agreement to participate was obtained from 10 members of the public with experience of NHS, local authority or commercial weight management programmes in the UK, 10 academics/policy makers/commissioners working in weight management, 10 weight management staff involved in delivering a lifestyle weight management programme for adults (without significant policy involvement), and 10 primary care staff with experience of referring patients to weight management programmes (Table S2).

With regard to members of the public, in line with the sampling framework, 6 of 10 had experience of commercial BWMIs (60%), 6 of 10 were of working age (60%) and 4 of 10 were male (40%) (Table S2). The 10 members of the public represented 9 different UK counties (6 Scottish counties and 3 English counties).

As per the sampling framework, 9 of the 10 academics/policy makers/commissioners were from England (90%), 4 of the 10 were academics (40%), 3 of the 10 were policy makers (30%) and 3 of the 10 were commissioners (30%) (Table S2).

Seven of the 10 primary care staff (70%) and 8 of the 10 weight management staff (80%) selected were from England (Table S2).

For the second Delphi process (stage 2, instrument/definition selection), 20 academics/policy makers/commissioners and 20 weight management staff were invited to participate and included those who had successfully completed all 3 rounds of the stage 1 Delphi. The stage 2 Delphi involved reading papers, looking at metrics and assessing validity of instruments/questionnaires. With such a level of knowledge and expertise required, members of the public and primary care staff were not involved in this stage of the Delphi process.

Broadly in keeping with our sampling framework, 16 of the 20 stage 2 academics/policy makers/commissioners group members were from England (80%), 11 of the 20 were academics (55%), 4 of the 20 were policy makers (20%) and 5 of the 20 were commissioners (25%) (Table S3).

With regard to weight management staff, as per our sampling framework, 14 of the 20 group members were from England (70%) (Table S3).

The research team conducting the study consisted of a clinical trialist/obesity physician, a health psychologist/trialist in weight management and behaviour change, a public health researcher/specialist advisor to PHE Obesity Team, and a researcher in cardiometabolic medicine.

Delphi survey

Delphi methodology was used to gain consensus from the expert group. Two separate Delphi processes (stage 1 and stage 2) were conducted using an online questionnaire system (www.clinvivo.com). Each Delphi process ran over three sequential rounds with the same group of participants (Figure 1). For both the outcome selection and outcome measurement/outcome definition selection (stage 1 and stage 2) Delphi processes, those who completed a questionnaire in round 1 were eligible to participate in round 2, and those who completed round 2 were eligible to participate in round 3. In short, in order for the expert group to reach consensus, only those completing a given questionnaire were eligible to complete the subsequent questionnaire.

The stage 1, outcome selection Delphi asked each expert to score the importance of an outcome measure for use in BWMI outcome reporting. The scale ran from 1-9 with 1-3 indicating that the outcome was unimportant, 4-6 indicating that it was neither unimportant nor important ('unsure') and 7-9 indicating that it was important. During rounds 1 and 2, participants were also given the opportunity to suggest additional outcomes. All outcomes, excluding any rated unimportant by consensus (see 'Statistical analysis' section) and including any appropriate new outcomes, were carried forward to the subsequent round (Figure 1).

During the stage 2, definition/instrument selection Delphi, experts were asked to score the appropriateness of outcome definitions and instruments for measurement of outcomes. Again, this was done using a 1-9 scale with 1-3 indicating that the definition/instrument was inappropriate, 4-6 indicating that it was neither appropriate nor inappropriate ('unsure') and 7-9 indicating that it was appropriate. During rounds 1 and 2, participants were once more given the chance to suggest additional instruments/definitions. As for stage 1, all instruments/definitions, excluding any rated unimportant by consensus (see ‘Statistical analysis’ section) and including any new instruments/definitions, were carried forward to the subsequent round (Figure 1).

For both stage 1 and stage 2 of the Delphi process, participant responses were summarised and fed back in subsequent rounds with participants receiving their own score and the expert group mean score for each outcome or instrument/definition.

Following round 3 of the stage 1 Delphi, consensus on the outcome set size and importance of outcomes was used to develop an outcome set. Similarly, following round 3 of the stage 2 Delphi process, a final instrument set matched to the COS was formed based on the consensus. In areas where there was no consensus, the study team adjudicated, taking account of free text comments.

Statistical analysis

As outlined in our published protocol[17], the Research ANd Development (RAND)/ University of California Los Angeles (UCLA) appropriateness method[21] was used to assess disagreement and importance/appropriateness (and thus define consensus). This involved calculating the mean score, the median score, the inter-percentile range (IPR, 30th and 70th), and the inter-percentile range adjusted for symmetry (IPRAS), for each item being rated. For a given item, disagreement was indicated when the ratio of IPR to IPRAS (the disagreement index) was greater than 1.

Importance/appropriateness was assessed simply as whether the mean and/or median rating fell between 1 to 3 (unimportant/inappropriate), 4 and 6 (unsure), or 7 and 9 (important/appropriate).

At the end of each Delphi round, the mean and median ratings were determined for individual outcomes/instruments and the distribution of ratings summarised (Figure 1). Free text comments were analysed qualitatively creating a narrative summary of responses based on the 9 domains used in the questionnaire.

RESULTS

Outcome selection

A list of 94 outcomes for reporting from BWMIs was generated from our review of the literature and systematic review process.

The 94 outcomes were mapped across appropriate domains by consensus of three members of the research team at a face to face meeting. The domains followed section headings used in the PHE SEF[9] and followed the weight management intervention chronological pathway (the order in which a BWMI would record outcome data as individuals progressed through the programme). There were 9 domains in total (Demographics, Physical Measurements, Physical Activity, Diet, Comorbidities, Lifestyle Behaviours, Psychological Factors, Programme Specific Outcomes and Length of Follow-up).

Delphi survey – stage 1/outcome selection

Round 1

The final list of domains and outcomes was used to develop an online outcome selection (stage 1) questionnaire. Within the questionnaire, an explanation/definition of each outcome was provided using lay terminology as identified by the research team and approved by Clinvivo staff. With the exception of the outcomes in the Demographics, Programme Specific Outcomes, and Length of Follow-up domains, all outcomes required measurement and reporting at both the first visit to a BWMI (baseline) and at the end of the programme/at follow-up. This resulted in a 148 item questionnaire with 75 outcomes for reporting at baseline and 73 outcomes at the end of the intervention. The stage

1, round 1, Delphi questionnaire can be seen, as it appeared to study participants, in Supporting Information 3. Of the 40 invited participants, 38 completed responses were received for the stage 1, round 1 Delphi questionnaire, representing a 95% response rate (100% of members of the public, academics, policy makers, commissioners and weight management staff, and 80% of primary care staff).

102 of 148 outcomes were rated as important by the expert group (median rating ≥ 7) with no evidence of disagreement between group members. The 102 outcomes rated as important were carried forward to the round 2 Delphi questionnaire (Table S4).

The remaining 46 outcomes were rated as being either unimportant or unsure (neither important nor unimportant) by the expert group (median rating ≤ 6.5, Table S4. For all but one outcome (1 month follow-up time point, disagreement index > 1), expert group members were again in agreement (Table S4). Outcomes rated as unimportant or unsure were not carried forward to round 2 (Table S4).

During the round 1 questionnaire, 19 additional outcomes were suggested by expert group members (Table S5 and Supporting Information 4). The study team decided that 4 of the 19 suggested outcomes were unique and valid, and would therefore be carried forward to the round 2 Delphi (Table S5), giving a total of 109 outcomes to be rated in this round (3 of the 4 additional outcomes were to be rated for reporting at both first visit and end of programme).

Round 2

The stage 1, round 2 Delphi questionnaire can be seen, as it appeared to study participants, in Supporting Information 5.

33/38 completed questionnaires were received, representing an 86.8% response rate (100% of academics, policy makers and commissioners, 90% of members of the public and 62.5% of primary care staff).

Following analyses of round 2 questionnaires, 87 of 109 outcomes were found to have been rated as important by the expert group (median rating ≥7). The remaining 22 outcomes were rated as unsure (median rating ≤ 6.5). No outcomes were rated as being unimportant and no disagreement was evident between group members for any of the ratings (Table S4). Participants' free text comments from round 2 can be seen in Supporting Information 6. No additional outcomes were suggested during this round.

In order to enable development of an outcome set of a manageable/practical size, the study team decided that outcomes would be split into three categories ('core', 'optional' and 'for exclusion') based on both their mean and median rating.

The 14 outcomes rated as most important with a mean rating >7 and a median rating ≥8 were designated as core for measurement and reporting by BWMIs (Table 1A). Of these 14 outcomes, 4 were to be measured and reported at both first visit and at the end of the programme. An additional 5 outcomes ('gender', 'ethnicity', 'deprivation category', 'learning disability' and 'physical disability') were then added to the core category. While these additional outcomes were rated as being important by the expert group, mean scores were not >7 and/or median scores were not ≥8. However, these outcomes are considered protected characteristics[22] and therefore should be reported in government-funded projects. Finally, an entirely new outcome, 'formally diagnosed with a mental health condition', was added to the core category as it was felt that its inclusion was necessary to ensure both a comprehensive COS and alignment with PHE key performance indicators[20]. Therefore, the core set included 20 outcomes for measurement and reporting by BWMIs (Table 1A).

Twenty two outcomes were rated as being reasonably important with a mean rating ≥6.5 and ≤7.1, and a median rating ≤8. These outcomes were designated as being optional for measurement and reporting by BWMIs. Of these 22 outcomes, 9 were to be measured and reported at both first visit and at the end of the programme. Of note, for 4 of these 9 ('blood pressure', 'cardiovascular risk', 'self esteem' and 'self confidence'), the mean rating was slightly less than 6.5 for the first visit time point. However, with the corresponding end of programme/follow-up time point meeting the rating criteria for the optional list, it was felt that these 4 outcomes should be included in order to ensure the follow-up measurement was meaningful with a baseline value to compare it to. As such, the optional set included 22 outcomes for measurement and reporting by BWMIs (Table 1B).

The 37 outcomes rated as being least important by the expert panel (mean <6.5 and median ≤7) were grouped together in the 'for exclusion' category. These outcomes would not be recommended for measurement and reporting by BWMIs unless participants gave a convincing argument for their inclusion during the round 3 Delphi (Table 1C).

Round 3

The stage 1, round 3 Delphi questionnaire can be seen, as it appeared to participants, in Supporting Information 7.

Prior to commencing the questionnaire, it was explained to participants that the results of the first 2 rounds of Delphi questionnaires had allowed lists of outcomes which would be considered core and optional for reporting by BWMIs to be made. It was explained that a list of outcomes to be excluded had also been drafted and that we would not recommend these outcomes be measured by BWMIs. Participants were informed that this would not mean that a weight management service could not measure these excluded outcomes should they wish to, but that measuring and reporting the other outcomes should be considered a higher priority.

Participants were asked to study the lists and indicate whether they agreed with the findings of the expert panel. They were advised that should they disagree with the findings, they would have the opportunity to express their disagreement and make suggestions as to any changes they felt should be made. It was made clear that if a number of participants were to express similar opinions, the lists would be altered appropriately.

The 33 expert group members who completed the round 2 questionnaire were invited to participate in the round 3 Delphi. All 33 members completed questionnaires, representing a 100% response rate for round 3. With 33/40 participants completing all 3 rounds of the stage 1 Delphi process, the overall response rate for stage 1 was 82.5% (100% of academics, policy makers and commissioners, 90% of weight management staff and members of the public, and 50% of primary care staff).

Following our analyses of the completed round 3 questionnaires, 25 of 33 participants (75.8%) indicated that they were in agreement with the core and optional outcome sets. Comments from the 8 participants who were not in agreement are included within Supporting Information 8. Having given these comments due consideration, the study team were of the opinion that no changes were required to the core or optional outcome sets (Tables 1A and 1B) prior to the stage 2 (instrument selection) Delphi process.

As outlined in Table 1A, the final list of core outcomes included ‘weight’ (at baseline and follow-up), ‘completion’ (at follow-up), ‘attendance’ (at follow-up), ‘BMI’ (at baseline and follow-up), ‘diabetes status’ (at baseline and follow-up), ‘participant satisfaction’ (at follow-up), ‘cost effectiveness’ (at follow-up), ‘age’ (at baseline), ‘Quality of Life (QoL) score’ (at baseline and follow-up), ‘reason for dropout’ (at follow-up), ‘adverse events/unintended consequences’ (at follow-up), ‘referral to specialist services’ (at follow-up), ‘12 months’ and ‘24 months’ follow-up time points, and ‘gender’, ‘deprivation category’, ‘physical disability’, ‘learning disability’, ‘ethnicity’ and ‘formally diagnosed with a mental health condition’ (all at baseline).

The final list of optional outcomes included ‘depression’ (at baseline and follow-up), ‘repeat referrals’ (at follow-up), ‘high blood pressure’ (at baseline and follow-up), ‘high future risk of diabetes’ (at baseline and follow-up), ‘overall measure of comorbidity’ (at baseline and follow-up), ‘binge eating disorder’ (at baseline and follow-up), ‘representativeness’ (at follow-up), ‘referral to linked services’ (at follow-up), ‘mobility issues’ (at baseline), ‘cardiovascular risk’ (at baseline and follow-up), ‘self confidence’ (at baseline and follow-up), ‘sources of referral’ (at follow-up), ‘prescription of anti-obesity medication’ (at follow-up), ‘high cholesterol/lipids’ (at baseline), ‘importance of weight loss’ (at baseline), ‘disordered eating’ (at baseline), ‘blood pressure’ (at baseline and follow-up), ‘self esteem’ (at baseline and follow-up), ‘reach’ (at follow-up) and ‘6 months’, ’18 months’ and ‘3 months’ follow-up time points (Table 1B).

With regard to outcomes for exclusion, 22 of 33 participants (66.7%) indicated that they were in agreement. Comments from the 11 participants who were not in agreement are included within Supporting Information 8. Again, following due consideration, the study team decided that no excluded outcomes should be retained/added to the optional outcome list prior to the stage 2 Delphi. The final list of outcomes for exclusion following the stage 1 Delphi process was, therefore, as outlined in Table 1C.

Outcome measurement instrument selection

By reviewing the trials identified by Hartman Boyce *et al.*[7] and our update, definitions and instruments that could be used for measurement of the core and optional outcomes selected during the stage 1 Delphi process were listed (Table S6). Further suitable definitions and instruments for these outcomes were added based on the study team’s knowledge (Table S6).

For simplification, outcomes for which the definition or instrument was well established or where only a single possible option was available were not included in the stage 2 process, while some outcomes within the optional outcomes set were combined; 'Binge Eating Disorder' was combined with 'Disordered Eating', and, although slightly different concepts, 'Self Esteem' and 'Self Confidence' were combined. Furthermore, an outcome relating to the presentation of results was added to the core set for inclusion in the stage 2 Delphi. Due to having specific instruments for their measurement, 'Learning Disability QoL Score' and 'Physical Disability QoL Score' outcomes were also included in the core set. In addition, as it had been borderline for inclusion based on rank, required only a yes/no answer with no patient burden and was specifically mentioned in NICE guidance[1] as a question for future research, the 'Repeat Referrals' outcome (mean rating of 7.1 and median rating of 7) was moved from the optional to the core outcomes list (Table S6).

Delphi survey – stage 2/outcome measurement instrument selection

Round 1

The stage 2, round 1 Delphi questionnaire can be seen, as it appeared to study participants, in Supporting Information 9. Documents 1-8 referred to within the questionnaire were provided in parallel and included full descriptions of all instruments and, where possible, peer-reviewed publications regarding their validity[23-26].

33/40 completed questionnaires were received, representing an 82.5% response rate (85% of weight management staff, 82% of academics, 80% of commissioners and 75% of policy makers).

Following analyses of completed questionnaires, 56 of 163 definitions/instruments were found to have been deemed appropriate by the expert group (median rating ≥ 7) with no evidence of disagreement between expert panel members (Table 2). The remaining 107 definitions/instruments were rated as unsure (neither appropriate nor inappropriate) by the expert group (median rating ≤ 6.5). The expert group were in agreement (disagreement index < 1.0) for 104 of these 107 items (Table 2).

For all but 8 outcomes, round 1 scores allowed discrimination between the definition/instrument options provided. In the majority of instances, options were selected for reporting if they were rated as important (median score ≥7). For outcomes where none of the definition/instrument options were rated as important (‘Learning Disability QoL Score’, ‘High Cholesterol/Lipids’, ‘High Future Risk of Diabetes’ and ‘Self-confidence and Self-esteem’), the highest scoring of the options deemed unsure were selected (Table 2). In cases where one of many definition/instrument options for an outcome received a much higher rating than the others, this option was selected for reporting and the lower scoring options were discarded despite some being rated as important (median ≥7). An example of this can be seen for the ‘Attendance’ outcome where item 11.1, ‘mean % of core/mandatory sessions attended by participants’ (median value of 8 and mean value of 7.9) was selected for reporting and items 11.3, ‘% of participants attending ≥80% of core/mandatory sessions’, and 11.4, ‘% of participants attending ≥70% of core/mandatory sessions’, (median values of 7 and mean values of 6.8 and 6.5 respectively) were discarded. Conversely, for the ‘Representativeness’ outcome, item 28.7, ‘based on other criteria’ was included for reporting despite being rated as unsure (median value of 5). This was because this item requested suggestions for additional measures and one of the free text suggestions provided (geographical location) was deemed suitable for reporting. Participants’ free text comments from round 1 can be seen in Supporting Information 10. Thirty five definitions/instruments relating to the 8 outcomes listed above were carried forward to the round 2 Delphi questionnaire (Table 2).

Round 2

The stage 2, round 2 Delphi questionnaire can be seen, as it appeared to study participants, in Supporting Information 11. Within this questionnaire, participants were required, for each of the 8 included outcomes, to rank the options provided in terms of their appropriateness for use or to select a single preferred definition/instrument. As stated, 35 definitions/instruments were carried forward from the stage 2, round 1 questionnaire. However, participants were asked to consider 31 options during the stage 2 questionnaire, the result of baseline and follow-up time points being combined where possible, and the addition of options representing a combination of definitions/instruments for a given outcome (Supporting Information 11).

The 33 expert group members who completed the stage 2, round 1 questionnaire were invited to participate in round 2 and 29/33 completed questionnaires were received, representing an 88% response rate (100% of weight management staff, 88.9% of academics, 66.7% of policy makers and 50% of commissioners).

As shown in Supporting Information 11, participants were asked to rank 7 definitions for measuring and reporting weight loss at follow-up in order of their appropriateness for use. Results are summarised in Table 3A. Based on mean and median ratings, all 4 potential definitions (items 3.1, 3.2, 3.3 and 3.4) were selected to be carried forward to the final definition/instrument selection Delphi (stage 2, round 3 questionnaire).

Similarly, the expert panel ranked 5 options pertaining to the presentation of results at follow-up in order of their appropriateness for use (Supporting Information 11). Results are shown in Table 3B. Based on mean and median ratings, item 7.5 (combining both items 7.2 and 7.3) was selected to be carried forward to round 3.

For the remaining 6 outcomes (‘Completion’, ‘Participant Satisfaction’, ‘Cost Effectiveness’, ‘Overall Measure of Comorbidity’, ‘Depression’ and ‘Importance of Weight Loss’), experts were instructed to select the most appropriate definition/instrument for measurement and reporting from the options provided (Supporting Information 11). Selection frequency for each option was determined and the option selected most frequently for a given outcome was then carried forward (Table 3C), the exceptions being ‘Participant Satisfaction’ and ‘Overall Measure of Comorbidity’. For the former, experts’ comments and scores indicated that neither of the suggested instruments (questionnaires) was ideal. Therefore, it was decided that both instrument options would be retained for round 3 but the expert panel would be informed that alternative methods to measure this outcome could be used. In the case of 'Overall Measure of Comorbidity', the majority of experts indicated that they had insufficient knowledge of the instruments and were therefore unable to select which would be most appropriate for use. Consequently, the most frequently selected of the remaining options, mean Edmonton Obesity Scale Score (EOSS) score, was selected to be carried forward to round 3.

Participants' free text comments from round 2 can be seen in Supporting Information 12.

Round 3

Experts were asked to study the final list of selected definitions/instruments and indicate whether they were in agreement with the findings of the expert panel. If participants disagreed with the findings they had the opportunity to express this disagreement and make suggestions as to any changes they felt should be made. It was made clear that should a number of experts express similar opinions, instruments/measurements would be altered appropriately. The stage 2, round 3 questionnaire is included, as it appeared to participants, as Supporting Information 13.

The 29 expert group members who completed the stage 2, round 2 questionnaire were invited to participate in the round 3 Delphi process and 27/29 completed round 3 questionnaires were received, representing a 93% response rate for this round (100% of weight management staff, 100% of academics, 50% of policy makers and 50% of commissioners). With 27/40 participants completing all 3 rounds of the stage 2 Delphi process, the overall response rate for stage 2 was 67.5% (85% of weight management staff, 72.7% of academics, 25% of policy makers and 20% of commissioners).

Following analyses of round 3 questionnaires, results revealed that 19/27 experts (70%) approved the results as presented and 8/27 experts (30%) did not. With regard to expert panel subgroups, 7/8 academics (88%) approved the results as presented and 1/8 (13%) did not. The participant who identified as a commissioner accepted the results as presented, as did the participant who identified as a policy maker. Of the weight management staff, 10/17 (59%) agreed with the results as presented and 7/17 (41%) did not. Therefore, the most disagreement and, consequently, free text comments came from weight management staff who tended to pre-empt their responses by stating that they partially accepted the results rather than rejecting them outright (Supporting Information 14). Comments suggested that the main concern was related to measures of diabetes status with participants questioning whether there was capacity in services to perform the necessary medical tests, who would fund these tests and whether performing them would place an unreasonable burden on weight management staff (Supporting Information 14). However, with the vast majority of the expert group in agreement with the results and free text comments of those not in agreement failing to provide a convincing argument for alteration of the final definition/instrument list, our core and optional outcome and definition/instrument sets were finalised and are included as Table 4. As shown, 'outcomes' within both sets were designated as being either process outcomes, outcomes or guidance for presentation of results (Table 4).

DISCUSSION

A COS is an agreed minimum set of outcomes for measuring and reporting for a specific area of health. Core outcome sets have been developed across a range of health areas, including bariatric and metabolic surgery[27]. While a recent study obtained expert panel consensus on recommendations for standard baseline assessment in medical obesity management clinics[28], to our knowledge, the study described herein is the first of its kind to develop a COS and corresponding definition/instrument set for BWMIs for adults with overweight and obesity. This is much needed in order to standardise reporting which, in turn, will lead to a better evidence base and improvements in weight management provision. Indeed, within the UK, PHE and Health Scotland have agreed to use this work to inform evaluation plans for adult BWMIs.

A wide range of sources, including the research literature and guideline and policy documents were used to generate lists of potential outcomes and definitions/instruments. Consensus as to which of these should be included in the final outcome sets was then determined by a group of individuals with wide-ranging expertise in behavioural weight management. This was achieved by means of the internationally-recognised Delphi process. Experts included members of the public with experience of BWMIs, academics/commissioners/policy makers working in weight management, weight management staff and primary care staff (referrers). There is no published agreement on the optimal size of an expert group[29]; pragmatism is required while ensuring a range of opinions is garnered. For this study, experts were selected according to our sampling framework to ensure they were representative of the UK as a whole, and the online nature of the Delphi process ensured that opinions expressed by members of the public were given equal weighting to those expressed by professionals. However, throughout the majority of the Delphi process, experts from each of the 4 groups were observed to be in agreement as to the importance of outcomes for reporting from BWMIs and the appropriateness of definitions/instruments for their measurement. In addition, retention rates for our experts were high throughout the Delphi process with 82.5% completing stage 1 (outcome selection) and 67.5% completing stage 2 (instrument selection). These high retention rates can be attributed to the nature of our recruitment and selection processes. In order to select a panel based on our sampling framework, potential experts were asked to provide information on geographical location etc. Those responding appropriately in a timely manner demonstrated their willingness to participate and their commitment to the process, and were therefore considered for Delphi expert panel selection. Those failing to respond to our requests were deemed unlikely to fully engage with the Delphi process and were not included in the selection process.

Experts agreed on a final core outcome and corresponding definition/instrument set consisting of 24 items which were designated as either processes, outcomes or guidance for presentation of results. As we may have expected, weight, body mass index, attendance, completion and cost effectiveness featured in the final COS and follow-up time points of 12 and 24 months were stipulated. Experts also agreed that an additional optional core outcome set was necessary. This included 19 items, again designated as either processes, outcomes or guidance for presentation of results, which BWMIs could report should they wish to do so. Both the core and optional outcome sets were observed to include outcomes relevant to patients, clinicians and commissioners/policy makers, reflecting the composition of our expert group.

While the vast majority of experts were in agreement with the final outcome and corresponding definition/instrument sets, some issues were raised by weight management staff with regard to the feasibility of the outcomes. With these concerns in mind, it should be noted that the measurement of each outcome is not considered mandatory for every patient/participant; the outcome sets are merely intended to serve as a guide for planned evaluations. A lack of funding and requirement for evaluation is a key issue for real-world services. The majority of outcomes in the core outcome set are generally measured during routine care but it is recognised that certain outcomes will prove more challenging for weight management staff, an example being the determination of haemoglobin A1c (HbA1c) levels if linking to routinely measured test results is not possible. In addition, information on longer term outcomes (at 12 months and 24 months) is likely to be difficult to obtain given the relatively short duration of the majority of BWMIs. Furthermore, those participants who regain weight are less likely to provide weight details or return to be weighed at a later stage. As such, research is needed in order to improve linkage to health records and to determine how best to persuade patients/participants to engage with longer term outcomes[1], perhaps by digital means, such as blue tooth scales or mobile apps. There is also a need for commissioners to consider the benefits of evaluation at the point of commissioning a service and ensuring that the service is funded sufficiently in order to gain meaningful insights[30].

This study was, of course, restricted to the UK. This is due to BWMIs and their settings within health services being fairly country-specific. For example, in France and the Netherlands there is no health insurance funding of BWMIs and, in the United States of America (USA), obesity services are tertiary, combining behavioural programmes with medication and bariatric surgery. Instruments can also be country-specific due to differences in language and health economic models, for example. In addition, ‘international’ studies are often tokenistic, including only a small percentage of participants from outside the country in which the study is set. Within the ‘international’ BARIACT study for example, the vast majority of professionals (95.2%) and patients (95.6%) participating were from the UK[27]. Our preference was to develop a core outcome set with a balanced stakeholder group using a sampling framework to ensure wide representation; to do this on a truly international scale would be impossible. Consequently, if used in an international context for trials or real world services, our core outcome and definition/instrument set may require further adaptation. Therefore, the next step may be to undertake international validation of the COS. This could involve consensus meetings with professionals and patients in other countries.

In conclusion, this study has used internationally recognised methodology to develop a COS for BWMIs. Its widespread adoption by both clinical trialists and weight management programmes will improve the quality of data from research studies and real-life services, thus improving the evidence base and weight management provision.

REFERENCES

(1) National Institute for Health and Care Excellence. Weight Management: lifestyle services for overweight or obese adults (PH53). 2014.

(2) Scottish Intercollegiate Guidelines Network. Management of Obesity: A national clinical guideline. February, 2010.

(3) Loveman E, Frampton GK, Shepherd J et al. The clinical effectiveness and cost-effectiveness of long-term weight management schemes for adults: a systematic review. *Health Technol Assess* 2011;15(2):1-182.

(4) Read S, Logue J. Variations in weight management services in Scotland: a national survey of weight management provision. *J Public Health (Oxf)* 2016;38(3):e325-e335.

(5) Public Health England. National mapping of weight management services. December 10, 2015.

(6) Borek AJ, Abraham C, Greaves CJ, Tarrant M. Group-Based Diet and Physical Activity Weight-Loss Interventions: A Systematic Review and Meta-Analysis of Randomised Controlled Trials. *Appl Psychol Health Well Being* 2018;10(1):62-86.

(7) Hartmann-Boyce J, Johns D, Aveyard P et al. Managing overweight and obese adults: The clinical effectiveness of long-term weight management schemes for adults (Review 1a). February 11, 2013.

(8) Public Health England website. https://www.gov.uk/government/organisations/public-health-england. Accessed January 2019.

(9) Public Health England. Standard Evaluation Framework for weight management interventions. March, 2009.

(10) All-Party Parliamentary Group on Obesity. The current landscape of obesity services: a report from the All-Party Parliamentary Group on Obesity. May 15, 2018.

(11) Atlantis E, Kormas N, Samaras K et al. Clinical Obesity Services in Public Hospitals in Australia: a position statement based on expert consensus. *Clin Obes* 2018;8(3):203-210.

(12) Canadian Obesity Network. Report Card on Access to Obesity Treatments for Adults in Canada. July 18, 2017.

(13) Gasoyan H, Tajeu G, Halpern MT, Sarwer DB. Reasons for underutilization of bariatric surgery: The role of insurance benefit design. *Surg Obes Relat Dis* 2018.

(14) Welbourn R, Pournaras DJ, Dixon J et al. Bariatric Surgery Worldwide: Baseline Demographic Description and One-Year Outcomes from the Second IFSO Global Registry Report 2013-2015. *Obes Surg* 2018;28(2):313-322.

(15) Williamson PR, Altman DG, Bagley H et al. The COMET Handbook: version 1.0. *Trials* 2017;18(Suppl 3):280.

(16) HTA Stage 1 guidance notes. National Institute for Health Research website. https://www.nihr.ac.uk/documents/hta-stage-1-guidance-notes/11743. Published May 31 2019. Accessed August 2019.

(17) Mackenzie RM, Ells LJ, Simpson SA, Logue J. Development of a core outcome set for behavioural weight management programmes for adults with overweight and obesity: protocol for obtaining expert consensus using Delphi methodology. *BMJ Open* 2019;9(2):e025193.

(18) Kirkham JJ, Gorst S, Altman DG et al. Core Outcome Set-STAndards for Reporting: The COS-STAR Statement. *PLoS Med* 2016;13(10):e1002148.

(19) Adult weight management services: collect and record data. GOV.UK website. https://www.gov.uk/government/publications/adult-weight-management-services-collect-and-record-data. Published June 21, 2017. Accessed January 2019.

(20) Adult weight management: key performance indicators. GOV.UK website. https://www.gov.uk/government/publications/adult-weight-management-key-performance-indicators. Published November 1, 2017. Accessed January 2019.

(21) Fitch K, Bernstein SJ, Aguilar DM et al. The RAND/UCLA Appropriateness Method User's Manual. Santa Monica, CA: RAND; 2001.

(22) Protected characteristics. Equality and Human Rights Commission website. https://www.equalityhumanrights.com/en/equality-act/protected-characteristics. Published January 8, 2019. Accessed January 2019.

(23) Gibbons E, Hewitson P, Morley D, Jenkinson C, Fitzpatrick R. The Outcomes and Experiences Questionnaire: development and validation. *Patient Relat Outcome Meas* 2015;6:179-189.

(24) Kolotkin RL, Norquist JM, Crosby RD et al. One-year health-related quality of life outcomes in weight loss trial participants: comparison of three measures. *Health Qual Life Outcomes* 2009;7:53.

(25) Townsend-White C, Pham AN, Vassos MV. Review: a systematic review of quality of life measures for people with intellectual disabilities and challenging behaviours. *J Intellect Disabil Res* 2012;56(3):270-284.

(26) Von KM, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. *J Clin Epidemiol* 1992;45(2):197-203.

(27) Coulman KD, Hopkins J, Brookes ST et al. A Core Outcome Set for the Benefits and Adverse Events of Bariatric and Metabolic Surgery: The BARIACT Project. *PLoS Med* 2016;13(11):e1002187.

(28) Ramachandran D, Atlantis E, Markovic T, Hocking S, Gill T. Standard baseline data collections in obesity management clinics: A Delphi study with recommendations from an expert panel. *Clin Obes* 2019;e12301.

(29) Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs* 2000;32(4):1008-1015.

(30) Public Health England and NICE. A Guide to Delivering and Commissioning Tier 2 Adult Weight Management Services. June 2017.

Authors' contributions

RMM and JL drafted the manuscript. LJE and SAS critically reviewed the manuscript. RMM and JL finalised the manuscript.

Funding statement

This work was supported by the CSO of the Scottish Government Health Department, grant reference number CGA/17/08.

SAS was supported by a Medical Research Council (MRC) Strategic Award (MC-PC-13027, MC_UU_12017_14 and SPHSU14).

Table and figure legends

Figure 1. Schematic outlining the two stage Delphi study. In order to develop a core outcome set and definition/instrument set, Delphi methodology was used to gain consensus from expert groups. Two Delphis (stage 1 and stage 2) were carried out online over three rounds of questionnaires. The stage 1 Delphi focused on development of a core outcome set. The stage 2 Delphi focused on corresponding definition/instrument selection. PHE, Public Health England; SEF, standard evaluation framework; KPI, key performance indicator.

Table 1A. Outcomes to be considered core for measuring and reporting by behavioural weight management interventions (BWMIs). Outcomes rated by the expert panel as being most important with a mean rating >7 and a median rating ≥8 were designated as core for measurement and reporting by BWMIs. *Mean scores were not >7 and/or median scores were not ≥8 but outcomes are considered protected characteristics. **New outcome added to ensure a comprehensive core outcome set. BMI, body mass index; QoL, quality of life.

Table 1B. Outcomes to be considered optional for measuring and reporting by behavioural weight management interventions (BWMIs). Outcomes rated by the expert panel as being reasonably important with a mean rating ≥6.5 and ≤7.1, and a median rating ≤8 were designated as being optional for measurement and reporting by BWMIs. *Mean scores <6.5 for the first visit/baseline time point but corresponding follow-up time point scores meet rating criteria for the optional list. HbA1c, haemoglobin A1c.

Table 1C. Outcomes not recommended for measuring and reporting by behavioural weight management interventions (BWMIs). Outcomes rated by the expert panel as being least important with a mean rating <6.5 and a median rating ≤7 were designated as being 'for exclusion' and would therefore not be recommended for measurement and reporting by BWMIs, unless participants gave

a convincing argument for their recommendation during the round 3 Delphi. NAFLD, non-alcoholic fatty liver disease.

Table 2. Stage 2 (instrument selection), round 1 Delphi results. 56 of 163 definitions/instruments were rated as appropriate by the expert group (median rating ≥ 7) with no disagreement between experts. 107 definitions/instruments were rated as unsure (median rating ≤ 6.5). The expert group were in agreement (disagreement index < 1.0) for 104 of these 107 items. IPR, inter-percentile range: IPRAS, inter-percentile range adjusted for symmetry; BMI, body mass index; T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus; HbA1c, haemoglobin A1c; QoL, quality of life; EQ-5D-5L, EuroQol 5-level EQ-5D version; SF12, 12-Item Short Form Health Survey; SF36, 36-Item Short Form Health Survey; IWQOL-Lite, 31-Item Impact of Weight on Quality of Life; OWLQOL, Obesity and Weight-Loss Quality of Life; PWI-ID, Personal Wellbeing Index–Intellectual Disability; OEQ, Outcomes and Experiences Questionnaire; NHS, National Health Service; FFT, Friends and Family Test; PHE, Public Health England; CV, cardiovascular; CVD, cardiovascular disease; HDL, high-density lipoprotein; HDR, high diabetes risk; OGTT, oral glucose tolerance test; CCI, Charlson Comorbidity Index; EOSS, Edmonton Obesity Staging System; HADS, Hospital Anxiety and Depression Scale; PHQ-9, Patient Health Questionnaire-9; ICECAP-A, ICEpop CAPability measure for Adults; WEMWBS, Warwick-Edinburgh Mental Wellbeing Scale; DIET, Dieter’s Inventory of Eating Temptations; TFEQ, Three Factor Eating Questionnaire; EDEQ, Eating Disorder Examination Questionnaire; BES, Binge Eating Scale; QEWP, Questionnaire on Eating and Weight Patterns.

Table 3A. Central tendency and spread of ratings for stage 2 (instrument selection), round 2 Delphi items relating to the measuring and reporting of weight loss at follow-up. Participants were asked to rank 7 definitions for measuring and reporting weight loss at follow-up in order of their appropriateness for use. Based on mean and median ratings, all 4 potential definitions (items 3.1, 3.2, 3.3 and 3.4) were selected to be carried forward to the final definition/instrument selection Delphi (stage 2, round 3).

Table 3B. Central tendency and spread of ratings for stage 2 (instrument selection), round 2 Delphi items relating to the presentation of results at follow-up. Participants were asked to rank 5 options pertaining to the presentation of results at follow-up in order of their appropriateness for use. Based on mean and median ratings, 2 items (items 7.2 and 7.3) were selected to be carried forward to the final definition/instrument selection Delphi (stage 2, round 3).

Table 3C. Selection frequencies for remaining stage 2 (instrument selection), round 2 Delphi items. Participants were instructed to select the most appropriate definition/instrument for measurement and reporting from the options provided for each outcome. Selection frequency for each option was determined and the option selected most frequently retained for the stage 2, round 3 Delphi. *Participants' comments and scores indicated that neither of the suggested instruments was ideal. Therefore, no instrument was selected. These two options will be given as suggestions but other methods could be used.**The majority of participants indicated that they had insufficient knowledge of the instruments and were therefore unable to select which would be most appropriate for use. Consequently, the most frequently selected of the remaining options, mean EOSS score, was retained for the stage 2, round 3. SD, standard deviation; IQR, interquartile range; OEQ, Outcomes and Experiences Questionnaire; NHS, National Health Service; FFT, Friends and Family Test; PHE, Public Health England; CCI, Charlson Comorbidity Index; EOSS, Edmonton Obesity Staging System; HADS, Hospital Anxiety and Depression Scale; PHQ-9, Patient Health Questionnaire-9; DIET, Dieter's Inventory of Eating Temptations.

Table 4. Core and optional outcome and definition/instrument sets. The expert group agreed on a final core outcome and corresponding definition/instrument set consisting of 24 items. 12 of these items were designated as processes, 8 were designated as outcomes and 4 were designated as guidance for presentation of results. Experts agreed on an optional outcome set consisting of 19 items; 5 processes, 11 outcomes and 3 items relating to presentation of results.