Supplementary Tables – Appendix

Table 1. Patient-reported outcome measures collected in the CHESS feasibility trial (All measures were self-completed at baseline, 2-weeks and 12-weeks).

Patient-reported Outcome Measure				
	Conceptual focus	**Response options / recall period**	**How to score/ interpretation**	**Key reference**
Headache-specific				
Chronic Headache Quality of Life Questionnaire (CHQLQ) Modification of the measure involved replacing the term 'migraine' with 'chronic headache' in the item stem. *These changes were supported by the developers (e-mail correspondence: Chris Bell, GSK, January 2016)*	Headache-specific quality of life. 14-items assess the functional impact of headaches across 3 domains: Role restrictions (RR) – restrictions to daily activities caused by headaches Role prevention (RP)(items 8-11) - prevented from engaging in daily activities due to headaches Emotional function (EF)(items 12-14) – the impact of headaches on the individuals emotional well-being	Six descriptive response options, ranging from 'None of the time' (1 point) to 'All of the time' (6 points); 4-week recall period Trial-specific modification to the MQLQ (v2.1) – the word 'migraines' was replaced with 'headaches'.	Domain item responses are summed: Role restrictions (RR)(items 1-7) Role prevention (RP)(items 8-11) Emotional function (EF)(items 12-14) Domain scores rescaled to a 0-100, where higher scores indicate better headache-related quality of life	*Martin, B.C., et al., Validity and reliability of the migraine-specific quality of life questionnaire (MSQ Version 2.1). Headache, 2000. 40(3): p. 204-15.*
Comparator measures: Headache-specific				
Headache Impact Test (HIT-6) .	A 6-item measure which purports to provide an overall assessment of headache impact on an individual's ability to function	Each item has five descriptive response options, with each awarded a specific number of points: 'Never' (6 points), 'Rarely' (8 points), 'Sometimes' (10 points), 'Very often' (11 points) and 'Always' (13 points). No recall period items 1-3 4-week recall period items 4-6	The score is the sum of item (points) responses. Index score ranges 36 to 78, with lower scores indicating better health / less impact on life Interpretative guidance: Scores ≤49 indicates little to no impact on life; Scores 50-55 indicates some impact on life;	*Kosinski, M., et al., A six-item short-form survey for measuring headache impact: the HIT-6. Qual Life Res, 2003. 12(8): p. 963-74*

			Scores 56-59 indicates substantial impact on life; Scores ≥60 indicates very severe impact on life.	
Comparator measures: Generic				
Short Form 12-item Health Survey version 2 (SF-12v2) *Website: https://www.optum.com/solutions/life-sciences/answer-research/patient-insights/sf-health-surveys/sf-12v2-health-survey.html*	A 12-item, non-preference based measure of generic health status, derived from the SF-36 [Ware 2002]. It assesses health across eight domains including physical and social functioning, and mental health.	Each item has between three and five descriptive response options. 4-week recall period	Item scores are transformed and standardised to compute two summary scales: physical component scale (PSC) and mental component scale (MCS) Scores are based on general population values (range from 0 (substantial limitation) to 100 (no limitation), standardised to have a mean of 50 (SD 10)	*Ware J Jr, Kosinski M, Keller SD. A 12-item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996;34:220–233. Jenkinson C, Stewart-Brown S, Petersen S, Paice C. Assessment of the SF-36 version 2 in the United Kingdom. J Epidemiol Community Health. 1999;53:46–50.*
EuroQoL EQ-5D-5L *Website: https://www.euroqol.org/*	Preference-based, generic measure of quality of life. Includes five items / domain descriptive system: mobility, self-care, usual activities, pain and discomfort, anxiety and depression.	Five response options per item (no problems / slight problems or moderate problems / severe problems or unable to do (or extreme pain or extreme anxiety/depression)). Health status on the day of completion	Simple reporting of item level scores can provide a simple description of health status. More usually responses are used to calculate a utility index score ranging -0.59 to 1.0, where 1.0 is perfect health and a score less than zero is considered a state worse than death [Kind et al. 1997]. Utility tariffs generated from a representative sample of the UK adult population (≥ 18years of age) (Kind, Dolan et al. 1998) were used to derive the utility weights for this study. UK general population normative values for the EQ-5D-3L 0.86 (SD 0.23)	*The EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. Health Policy. 1990;16:199–208. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20:1727–1736.*

EuroQoL EQ-Visual Analogue Scale (EQ-VAS)	The second component of the EuroQoL – single items measure of general health	Single 10cm vertical (thermometer) on which respondents rate their overall health 'today'	Overall health today from 0 (worst imaginable) to 100 (best imaginable). UK general population normative values 82.48 (SD19.96)	
Comparator measures: Domain-specific				
Hospital Anxiety and Depression Scale (HADS)	A 14-item measure of anxiety and depression.	Two domains each consist of seven items, with four-point, descriptive response options ranging 0 to 3. Recall period 'past week'	Domain item responses are summed: Anxiety - items 1, 3, 5, 7, 9, 11, 13 Depression - items 2, 4, 6,8,10, 12, 14 Domain scores range 0 to 21. Interpretative guidance: 0-7 'normal' 8-10 mild anxiety or depression 11-15 moderate anxiety or depression >/=16 severe anxiety or depression	*Zigmond, A. S., Snaith, R. P. The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scandinavica, 1983. 67(6): p.361–370.*
Pain Self-Efficacy Questionnaire (PSEQ)	A 10-item measure of an individual's confidence in performing a particular behaviour or task despite of their pain	Each item has seven possible response options, ranging from 'Not at all confident' (0 point) to 'Completely confident' (6 points). No recall period	Item scores are summed. Index score range 0 to 60, where the higher score reflects stronger self-efficacy beliefs	*Nicholas, M.K., The pain self-efficacy questionnaire: Taking pain into account. Eur J Pain, 2007. 11(2): p. 153-63.*
Social Integration Subscale of the Health Education Impact Questionnaire (heiQ)	A 5-item domain (one of eight) included in the HEiQ. The SIS assesses ability to integrate in society (The HEiQ assesses the impact of patient education programmes in chronic conditions)	Four response options: 'Strongly disagree' (1 point), 'Disagree' (2 points), 'Agree' (3 points) and 'Strongly agree' (4 points). No recall period	Item scores are summed. index score range 5 to 20, where higher scores indicate higher levels of social interaction, a higher sense of support, and confidence in seeking support from others. Lower scores suggest greater feelings of social isolation because of illness.	*Osborne, R., et al. The Health Education Impact Questionnaire (heiQ): an outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. Patient Educ Couns, 2007. 66(2): p.192-201.*

Headache-specific Health Transition Questions	Patient-reported health transition items detailing the size and direction of change in health over a specified period are widely used as patient-derived, external criterion for change. The two and 12-week questionnaires included questions which asked patients to rate, overall, if they felt that their headaches were: much better / better / the same / worse / much worse on a 5-point scale.	*de Vet, C.W., Terwee, C.B., Mokkink, L.B., Knol, D.L. (2011) Measurement in Medicine: a practical guide. Practical Guides to Biostatistics and Epidemiology. Cambridge University Press*.

Table 2. Data analysis plan and interpretation

	Description	Analysis and interpretation
Data quality and measurement acceptability		
Completion rates	Item and scale level missing data reported as a reflection of measurement acceptability [6,7,15,16]	Item level and scale level score distribution and the percentage of computable scores reported
Item-total correlation (corrected) (cITC)	The extent to which items are adequate reflections of the common underlying latent construct [15,16]	Corrected item-total correlation. Values between +0.40 and +0.60 suggest moderate levels of inter-item correlation, supporting convergent validity; values greater than +0.70 suggest that there may be item redundancy [15,16]
***Interpretability* - t**he ability to assign qualitative meaning to a score or change in score (https://www.cosmin.nl/wp-content/uploads/COSMIN-definitions-domains-measurement-properties.pdf)[14,15]		
End-effects	Where more than 15% of respondents score the minimum (floor) or maximum (ceiling) score [6,15]	-
Minimal important change (MIC)	The MIC is defined as the smallest change in score perceived as important by participants) [14,15]	Calculated as the mean change score for people reporting 'minimal change' in headache at 12-weeks on the headache-specific health transition questionnaire (HTW)(that is, 'better' or 'worse').
Structural validity and internal consistency		
Structural validity	Structural validity, a component of construct validity, evaluates a measures underlying construction, the presence of sub-domains and item behaviour [6,14-16] Due to the CHQLQ item stem modification, an exploratory factor analysis (EFA) was conducted on baseline data, hypothesizing that the original three-factor solution would be retained. Confirmatory factor analysis (CFA) was then performed to confirm the proposed three- and one-domain factor structures of the CHQLQ and HIT-6, respectively.	***Exploratory factor analysis (EFA)*** Data was entered into a model with variamax rotation. Absolute item loadings ≥0.45 were accepted as sufficient correlation with a principal component to support domain inclusion. ***Confirmatory Factor Analysis*** Performed using lavaan [http://lavaan.ugent.be/tutorial/tutorial.pdf] in R version 3.3.1 (R Core Team, 2016) by maximum likelihood and full information maximum likelihood (FIML) for the missing data. The latent factors were standardised to have mean 0 and standard deviation of 1 to allow free estimation and easily interpretable factor loadings.Factor loadings exceeding 0.3-0.4 were judged to be meaningful [16,17]; the model fit was examined using Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). A CFI and TLI of >0.95 and a RMSEA of <0.05 were considered as adequate fit. For moderate fit, CFI and TLI values of >0.90 and RMSEA of <0.08 were used.
Internal consistency	Assesses the relationship (interrelatedness) between items within a measure (or sub-domains), reflecting the total number of items and their average correlation [15,16]	The internal consistency of the three CHQLQ domains and the HIT-6 - was assessed by calculation of Cronbach's alpha [15,16] Values between 0.7 and 0.90 suggest a good to excellent agreement between items and the total (domain) score [15,16]
Reliability and measurement error – the degree to which a measure is free from measurement error		

Test-retest reliability	The extent to which scores for patients who have not changed are the same for repeated assessments over time (temporal stability) [6,14-16]	Two-week test-retest reliability was assessed in patients who indicated on health transition item that their headaches had remained stable [15,16] The intra-class correlation coefficient (ICC 2,1) was used to measure the level of agreement between test and re-test.[15,16] For group comparisons, levels of reliability greater than 0.70 are recommended, with high levels (>0.90) required for individual-level assessment.
Measurement error - Standard Error of Measurement (SEM) - Smallest Detectable Change (SDC)	The systematic and random error of a patient's score that is not attribute to true changes in the construct to be measured (https://www.cosmin.nl/wp-content/uploads/COSMIN-definitions-domains-measurement-properties.pdf) The extent of absolute measurement error is determined by calculation of the Standard Error of Measurement (SEM). The SEM supports score interpretation by accounting for variability, or error, in measurement - only a change greater than measurement error is considered 'real'.[15,17] The Smallest Detectable Change (SDC) represents the smallest change in score that is greater than measurement error. The SDC allows one to rule out measurement error (i.e. distinguishing measurement error from true change) when assessing the reliability of a self-reported measure to detect change in health status. Thus, a score change greater than the SDC value is necessary to provide evidence of true change (improvement or deterioration) in health-status.	Standard Error of Measurement (SEM) Calculated using a two-way random effects model [6,15] The SEM was subsequently converted into the smallest detectable change (SDC). The SDC was calculated both for individuals and groups [6,15,19] - SDCindividual = SEM x 1.96 x √2 - SDCgroup = 1.96 x √2 x SEM √n (where n is the group size)
Construct and Content Validity *– the degree to which a measure measures what it purports to measure*		
Content Validity – qualitative evidence in support of purported measurement focus	The degree to which the content of the PROM is measures the construct(s) is purports to measure (https://www.cosmin.nl/wp-content/uploads/COSMIN-definitions-domains-measurement-properties.pdf [14-16] Evidence that details the clarity of measurement content in terms of relevance, comprehensiveness and comprehensibility with respect to the purported measurement focus (construct of interest) and the target population (for example, chronic headache)[14-16]	Semi-structured cognitive interviews were conducted with a purposive sample of patients with confirmed chronic headache to explore the relevance, acceptability, clarity and comprehensiveness of the measures, as per the four stages of cognitive processing: [21,22] - Comprehension: the process of making sense of the questions and developing a response - Memory retrieval: the process of accessing relevant information to enable a response - Judgement: the process to determine if memory retrieval is accurate and complete - Response mapping: the process by which an appropriate response is selected Several overarching questions sought to explore how patients determined an improvement in their headache, and if specific questions were missing. Interviews were continued until thematic saturation was achieved.

		Participants were interviewed within 24-hours of questionnaire self-completion. Verbal prompting was used to facilitate the interview process. To counteract fatigue-bias, the CHQLQ and HIT-6 were alternately completed. All interviews were audio recorded, transcribed verbatim, and checked for accuracy (VN). Framework analysis[23] and cross-case comparison was used to generate themes, informed by PROM item content, relevance and the additional over-arching questions. NVivo software was used to organise the data. Data was independently explored by two researchers (VN,KH). Emergent themes were discussed and interpreted with a third researcher (FG).
Construct Validity – quantitative evidence in support	The degree to which PROM scores are consistent with hypotheses, and based on the assumption that the PROM is a valid measure of the construct to be measured (https://www.cosmin.nl/wp-content/uploads/COSMIN-definitions-domains-measurement-properties.pdf) [14-16] Assessed by correlating the scores for separate measures to assess the convergent validity of related domains (Pearson's correlation coefficient): it was expected that related constructs would correlate more strongly.	Hypothesised theoretical associations between the three domains of the CHQLQ and comparator measures were considered a priori (Appendix Table 3). The RF and RP domains of the CHQLQ and HIT-6 measure related domains and hence a stronger association than with the EF domain was hypothesized. Similarly, the SF-12 PCS and EQ-5D-5L have a greater focus on physical aspects of health, and a stronger association with the RL and RP domains was hypothesized than with the EF domain; but a stronger association between the EF and the SF-12 MCS was hypothesized. A stronger association between the EF domain and the HADS-A and HADS-D was hypothesized. Several items within all three domains of the CHQLQ consider the social impact of headache, and hence moderate to strong association with the HEiQ SIS was hypothesized. The focus of the PSEQ is one's ability (and confidence) in managing pain and engaging in (largely) physical and social activities; hence a moderate association with the RL and RP domains, but small association with the EF domain was hypothesized.
Responsiveness *- the ability of a measure to detect real change in health over time that is greater than measurement error*		
- Smallest detectable change (SDC) in score	To understand the smallest change in score that is greater than measurement error in patients reporting change in headache at 12-weeks	***Standard Error of Measurement (SEM) and Smallest Detectable Change (SDC)***: To represent the smallest change in score that is greater than measurement error in those patients reporting change in headache in the headache-specific health transition question at 12-weeks, we calculated: - The absolute measurement error at 12-weeks (Standard Error of Measurement (SEM)) - The SEM was subsequently converted into the smallest detectable change (SDC) [14-16,19,20] - The SDC was calculated for both individuals and groups: - SDCindividual = SEM x 1.96 x √2 - SDCgroup = 1.96 x √2 x SEM √n (where n is the group size) ***Minimal important change (MIC):*** - calculated as the mean change in those who report a minimal improvement / deterioration on headache-specific health transition question at 12-weeks. ***Minimal important difference (MID):***

		- calculated as the mean change in score of those who were 'somewhat better' minus the mean change in those who were the same on headache-specific health transition question at 12-weeks.
- Criterion-based assessment of responsiveness	To understand the ability of measures to discriminate between patients whose headache had improved or deteriorated. This change was captured by the patient-reported headache-specific health transition question (an external criterion or 'gold standard of change).[6,14-16]	***External criterion:*** First, the level of correlation between change scores on CHQLQ and HIT-6 and the transition item was calculated; a level of agreement of 0.3 to 0.5 was considered acceptable as a marker of change.[24] ***ROC Analysis and Area Under the Curve (AUC)*** Receiver operating characteristic (ROC) curves were used to assess the ability of the measures to discriminate between patients who had improved or deteriorated, as per patient-self report of change in headache status at 12-weeks. Respondents were dichotomized in three ways: - i) headache 'much better' versus headache 'better, about the same or worse'; - ii) headache 'much better or better' (that is, the improved group) versus headache 'the same or worse' (the not improved group); and - iii) headache had 'improved or remained about the same' versus headaches had 'deteriorated'. The larger the area under the curve (AUC) (on a scale of 0.5 (no discriminatory power) to 1.0 (perfect discrimination)), the more sensitive the measure at detecting differences in the external indicator. An AUC of > 0.70 is considered as sufficiently discriminatory[6,15,16], whilst an AUC of 0.5 suggests no discriminatory power.
- Effect size (ES) and Standardised Response Mean (SRM) statistics	Distribution-based assessment (longitudinal validity): a priori defined hypotheses about the expected magnitude of differences in changes between defined groups (defined by responses to the 12-week headache-specific transition item) were proposed and tested.[15] Effect size classification[15]: Small 0.20 Moderate 0.50 Large 0.80	Effect Size (ES): mean change divided by baseline SD) Standardised Response Mean (SRM): mean change divided by SD of the change score Both values were calculated for sub-groups of patients in each health transition category. Given that this was a feasibility study with patients not receiving any intervention and the follow-up period was short, the main hypotheses to be tested for responsiveness were: • ES and SRM would be <0.2 for patients who reported no change in headache • ES and SRM would be >0.2 for patients reporting slight improvement • ES and SRM would be >0.5 for patients reporting improvement (much better) • ES and SRM would be greater for patients indicating an improvement in their headache than those indicating no change.

Table 3. Convergent validity matrix: hypothesized associations (size and direction) between CHQLQ and comparator measures

	Headache-specific		Generic				Domain-specific			
			Profile		Single item – General	Utility	Emotional well-being			
	Headache-specific QL	**Impact**	**Physical function**	**Mental well-being**		**Health Status**	**Anxiety**	**Depression**	**Pain self-efficacy**	**Social Integration**
CHQLQ	**CHQLQ**	**HIT-6**	**SF-12 PCS**	**SF-12 MCS**	**EQ-VAS**	**EQ-5D-5L**	**HADS - A**	**HAS - D**	**PSEQ**	**SIS-HEiQ**
Role Restriction (7 items)	Strong, positive with RP; Moderate positive with EF	Strong, positive	Moderate to strong, negative	Small to moderate, negative	Moderate, positive	Moderate* (to strong), positive	Small to moderate, positive	Small to moderate, positive	Moderate, positive	Moderate, positive
Role Prevention (4 items)	Strong, positive with RR; Moderate positive with EF	Strong, positive	Moderate to strong, negative	Small to moderate, negative	Moderate, positive	Moderate* (to strong), positive	Small to moderate, positive	Small to moderate, positive	Moderate, positive	Moderate, positive
Emotional Function (3 items)	Moderate, positive with RR and RP	Moderate to strong, positive	Small, negative	Moderate to strong, negative	Small to moderate, positive	Moderate, positive	Moderate to strong, positive	Moderate to strong, positive	Moderate, positive	Moderate, positive

Footnote: Strength of association (Cohen): small <0.30; moderate 0.31 to 0.69; strong >0.70. *EQ-D item content – stronger focus on physical function (mobility, usual activities, self-care), so stronger association with physical than with emotional domains hypothesized