EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

12 April 2018
EMA/CHMP/SAWP/226829/2018
Product Development Scientific Support Department

Qualification Opinion on Proactive in COPD

Draft agreed by Scientific Advice Working Party	26 October 2017
Adopted by CHMP for release for consultation	09 November 2017[1]
Start of public consultation	20 December 2017[2]
End of consultation (deadline for comments)	29 January 2018[3]
Adoption by CHMP	22 March 2018

Keywords	Activity monitor, chronic obstructive pulmonary disease, clinical trial, COPD, endpoint, patient reported outcome, physical activity, PRO.

[1] Last day of relevant Committee meeting.
[2] Date of publication on the EMA public website.
[3] Last day of the month concerned.

30 Churchill Place ● Canary Wharf ● London E14 5EU ● United Kingdom
Telephone +44 (0)20 3660 6000 **Facsimile** +44 (0)20 3660 5555
Send a question via our website www.ema.europa.eu/contact

An agency of the European Union

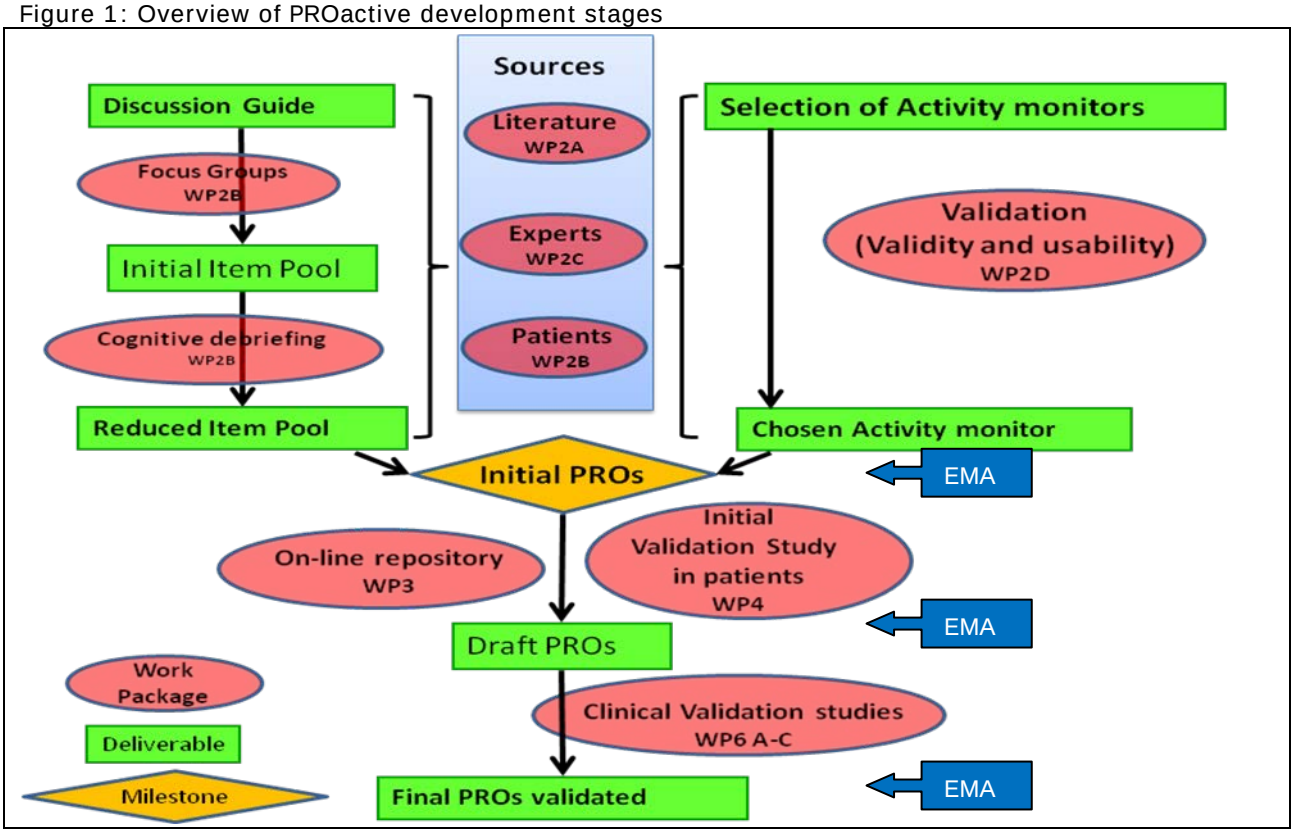
© European Medicines Agency, 2018. Reproduction is authorised provided the source is acknowledged.

Background information based on the Applicant's submission

Under the Innovative Medicines Initiative Joint-Undertaking (IMI-JU) framework, the public-private PROactive Consortium developed two Patient Reported Outcome (PRO) instruments to capture physical activity (PA) data in patients with Chronic Obstructive Pulmonary Disease (COPD) in clinical trial settings. One of those tools is the D-PPAC which is supposed to enable daily data collection (recall period of 1 day). The other developed PRO tool is the C-PPAC with a recall period of 7 days, intended to collect PA data during specified clinical study visits. The two PRO instruments have been developed as 'hybrid' tools, i.e. classical questionnaire items are combined with activity monitor readouts collected separately. The Consortium has produced electronic and paper-pencil versions of both the D-PPAC and C-PPAC instruments. Also, translations to several languages have been done for both tools. The English versions of the D-PPAC and the C-PPAC can be found in [1, 2].

During the development/validation phase, the Consortium sought advice from EMA in 2011 and in 2013 via the qualification advice procedure. These advice requests introduced the project, described the proposed conceptual framework (CFW) and sought advice on elements of the Consortium's approach to develop and validate the PRO instruments. In the framework of this (now third) interaction with EMA, the Consortium presented validation work carried out in their project's last phase (work package 6, WP6), which was based on 'final' versions of the two PRO instruments. Figure 1 below illustrates the project's work flow and its structure consisting of three important work-packages (WP2, WP4 and WP6). Details on these work-packages as well as corresponding assessment comments are found in a later section of this document.

Figure 1: Overview of PROactive development stages



Based on the totality of validation work as presented, the Consortium suggests that the PRO tools are ready for use in clinical trial settings having similar COPD patient populations as chosen in the WP4/WP6 trials.

The disease/condition in which the PPAC instruments are intended to be applied

COPD, the 3rd leading cause of death worldwide, represents an important public health challenge that is both preventable and treatable. COPD is a major cause of chronic morbidity and mortality throughout the world; many people suffer from this disease for years, and die prematurely from it or

its complications. Globally, the COPD burden is projected to increase in coming decades because of continued exposure to COPD risk factors and aging of the population (GOLD 2015).

Physical inactivity and its associated symptoms as a consequence of COPD are a hallmark of the disease potentially contributing to the disease progression (Hopkinson and Polkey, 2010). Patients are discouraged from being physically active due to the complex interplay of impaired exercise tolerance, symptoms, exacerbations and co-morbidities (e.g. heart disease, osteoporosis, musculoskeletal disorders, and malignancies) which may also contribute to restrictions of activity. Impaired activity leads directly and indirectly to increased morbidity and even increases mortality in COPD. The PA in which patients engage is the net result of the capacity patients have available to engage in and their active choice to use the available capacity.

As a consequence, both disease impact, mainly determined by symptom burden and activity limitations, and future risk of disease progression (e.g. exacerbations) should be considered when managing patients with COPD (GOLD 2015).

Drug developers have traditionally used spirometry, laboratory parameters, exercise capacity, clinical events (e.g. exacerbations) and/or health related quality of life as clinical trial outcome measures, which do not fully cover the patients' experience of the consequences of the disease.

While it is important to measure changes in respiratory function and symptom endpoints when evaluating new treatments in COPD, measuring their impact on aspects of daily life such as PA may be more meaningful to patients and physicians/healthcare providers. There is now considerable evidence that the level of FEV_1 is a poor descriptor of disease status (GOLD 2015).

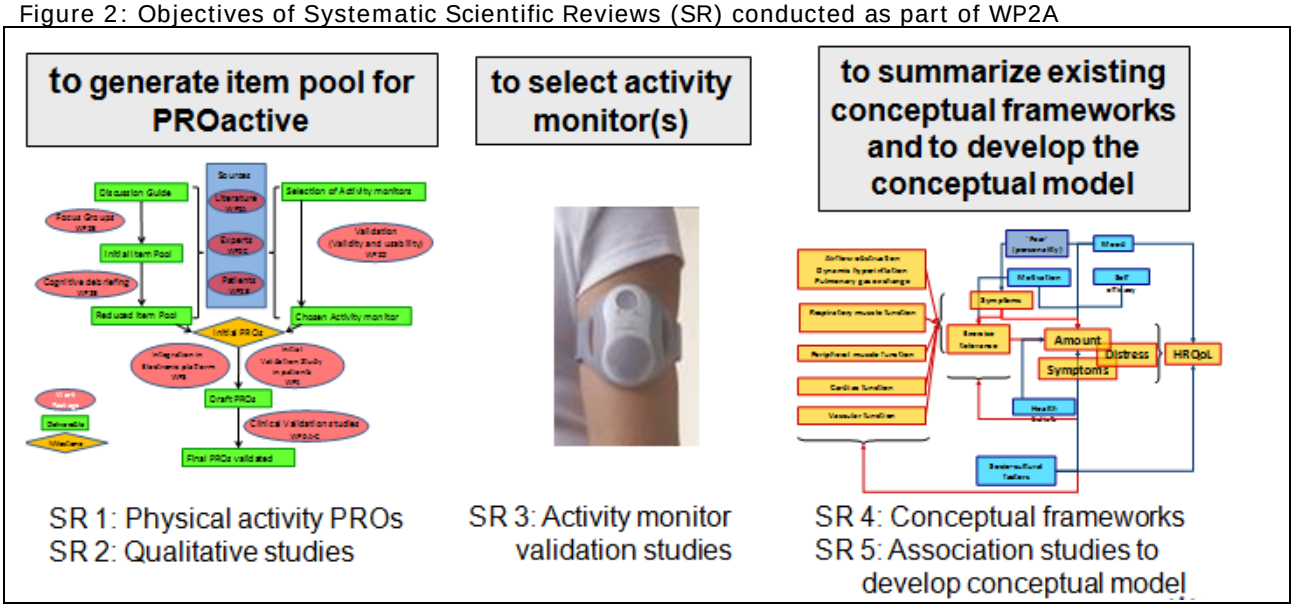
Physical activity as defined by Caspersen (any bodily movement that results in an increase in energy expenditure) can be measured with activity monitors (Caspersen et al. 1985). However these devices were, at the outset of the present project not well validated in COPD. More importantly they provide only quantitative indices of PA and do not capture the patient's experience with PA. A number of exercise capacity measures exist, e.g. field walking tests (Holland et al. 2014) or ergometry, which can inform researchers and developers about the patients' capacity for exercise. However, engagement in PA is a different concept, as not only it calls on the patient's physiological capacity, but also refers to a patient's self-efficacy and willingness to engage in activities. The latter two are potentially influenced by a complex and individual interplay of exercise related symptom perception, past behavior, health beliefs and motivation. Capturing all the dimensions of daily PA that are relevant to patients should provide a unique perspective of treatment effectiveness. However, despite its importance, no (other) existing PRO captures PA in a way that it maximally reflects the experience of patients with COPD. Also, there is no PRO that is sensitive enough to measure small but important changes in PA in clinical trials.

Presentation of development, validation and regulatory assessment of the PROs

Early development work forming the basis of both PRO instruments was carried out in the framework of Work Package 2 (WP2). There were 4 sub-work-packages that contributed to the development: systematic reviews of the literature (WP2A), patient input (WP2B), input from experts (WP2C) and the validation and selection of activity monitors (WP2D).

Under WP2A five systematic reviews of the literature have been done. Figure 2 illustrates the different objectives of these reviews.

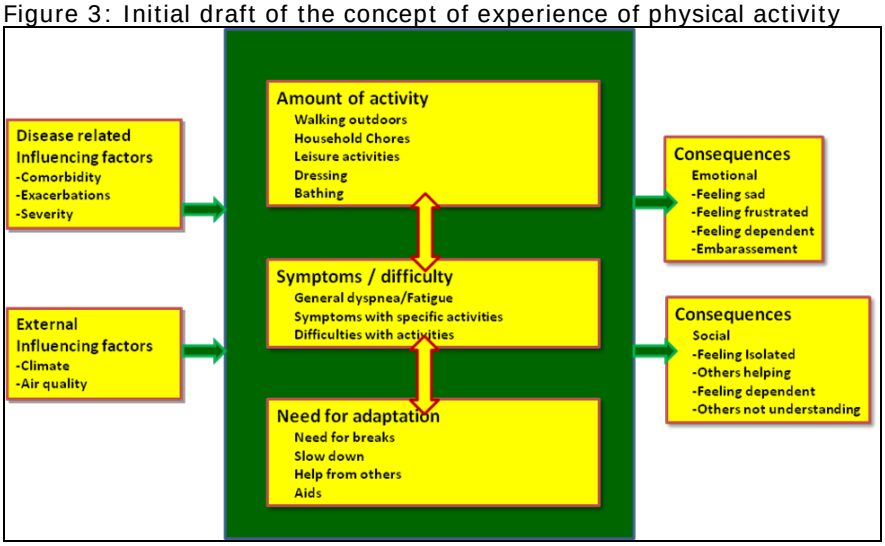
Figure 2: Objectives of Systematic Scientific Reviews (SR) conducted as part of WP2A



In summary, the literature reviews have helped to support the construct of the initial PROactive conceptual framework and the drafting of the endpoint model, developed specifically for patients with COPD, which is the intended population in which the PRO tools are supposed to be used. Reviews also revealed that no valid instruments or scales existed at the time of development start which would comprehensively capture PA from a COPD patient perspective. For more detailed descriptions of the outcome of the WP2A-reviews the reader is referred to [7, 8, 9].

In parallel to WP2A, another work package WP2B covered qualitative research involving COPD patients. This work package comprised one-to-one interviews, focus groups and cognitive debriefings which were conducted in four European countries: the UK, the Netherlands, Belgium and Greece. Involved COPD patients had different disease severity level. 116 patients participated in this qualitative research. WP2B activities allowed identification of the draft concept of experience of PA (figure 3).

Figure 3: Initial draft of the concept of experience of physical activity



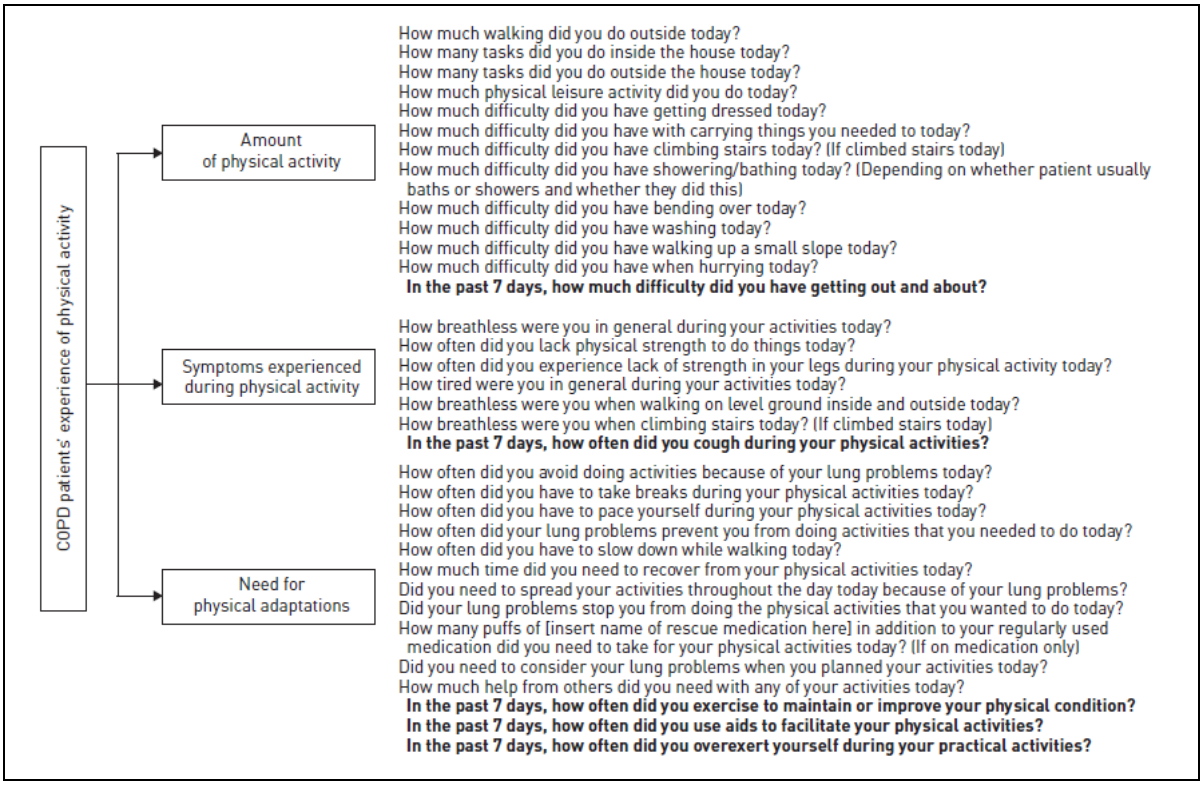
The qualitative studies also generated sufficient potential items shown to be of 'universal' importance to patients. An initial item pool was derived and items thereof were tested in WP4 in conjunction with the two selected activity monitors (see subsequent sections).

The work package WP2C assigned to expert input in the early stages of the instrument development was primarily implemented to determine the criteria to characterize the general patient population or give advice on the item pool. Here PRO developers used the complementarities of highly specialized experts in their respective fields from 18 different organizations actively involved in the PROactive

consortium. In addition, as part of the Advisory Board, the PROactive consortium has met bi-annually with a further set of 12 clinical and PROs experts as well as members from regulatory agencies that provide guidance on the PRO development and validation. Furthermore, through the European Respiratory Society, who was partner within the project, the consortium was also able to consult with multidisciplinary experts at key stages of the PROs development to ensure that the construct meets the clinicians' expectations.

Based on literature review, patient- and expert input, the initial conceptual framework (as shown in figure 4) was developed. Of note, items for the clinic visit PRO were similar to the items of the PRO to be completed on a daily basis, with the exception of the items in bold, which only appeared in the clinic visit PRO. This preliminary conceptual framework comprised 3 domains: 'Amount of PA', 'Symptoms experienced during PA' and 'Need for adaptations'.

Figure 4: Initial conceptual framework



This initial conceptual framework was subject to discussion during the first interaction with the SAWP qualification team (QT). In the course of assessing the first qualification advice request, the QT challenged the assumption that a PRO tool based on the domains 'symptoms during PA', 'amount of PA' and 'need for physical adaptations' will indeed be optimal to meet the Consortium's goal to have a reliable and valid measure for PA in COPD patients. Especially the 'symptoms'-domain was felt to contribute only little direct information about actual PA. At that time the Consortium explained the findings from qualitative interviews with a variety of patients with COPD, namely that symptoms patients experience during PA as well as adaptation required relate to the amount of PA they actually do. Although the QT agreed that all these themes related to PA are closely interlinked, and that the three proposed domains may be exhaustive to cover all relevant aspects to derive a PA score, it was considered important that the wording of the items (especially for the symptom-domain) reflects the link to (limited) PA. A pure domain on COPD symptoms without such a link was doubted to be supportive for a new concept. Concern was expressed that the new PRO tools would conceptually be very similar to already existing COPD questionnaires. In subsequent development and validation steps, the Consortium considered that point of criticism. The result was eventually an altered conceptual framework, not comprising a 'symptom'-domain anymore (see later sections).

One further aspect discussed with the Consortium at that stage of development was that improved PA should generally not be at the expense of other aspects of QOL in COPD patients. It was recommended

by the QT that this issue required dedicated investigations during PRO validation. The Consortium agreed and referred to their plans to also include measures of health status or health-related quality of life in the PROactive studies planned to investigate this issue. Furthermore, it was mentioned that most clinical studies in COPD include measures of health status or health-related quality of life, which would allow for investigating such a potential impact in specific drug developments later on.

In relation to the Consortium's goal to adequately cover the theme 'amount of PA' with their PROs, the idea of implementing read-outs from PA monitoring devices was introduced early during development. Early plans to possibly develop the PROs as hybrid tools merging monitor readout data with item response data were supported by the QT. PA monitors are frequently used to estimate levels of daily PA. A variety of PA monitors are available to measure bodily movement. These devices use piezoelectric accelerometers, which measure the body's acceleration, in one, two or three axes (uniaxial, biaxial or triaxial activity monitors). Signals are transformed into various measures of energy expenditure using specific algorithms, or are summarized as activity counts or vector magnitude units (reflecting acceleration). With the information obtained in the vertical plane or through pattern recognition, steps or walking time can also be derived by some monitors.

Reduced PA is an important feature of COPD. However, most of the monitors that were available at project start had been validated in healthy subjects, but not necessarily in patients with chronic diseases. As patients are less physically active and move slower than healthy subjects, the validity of these monitors to pick up movement needed to be evaluated further.

With work-package WP2D, two studies were conducted to identify suitable activity monitors to be used in validation studies as part of the PROactive instruments.

The first study, carried out in laboratory environment, followed the aim to evaluate the validity of six monitors in COPD patients (ranging in severity from mild to very severe according to GOLD stages) against a gold standard of indirect calorimetry in the form of VO2 data from a portable metabolic system. It was hypothesized that triaxial activity monitors (transducing body's acceleration in three axes) would be more valid tools when compared to uniaxial activity monitors. Indeed, the study found that three triaxial activity monitors (Dynaport Move Monitor, Actigraph GT3X and SenseWear Armband) were the best monitors to assess standardized and common physical activities in the range of intensity relevant to patients with COPD. Changes in walking speed were most accurately registered by the Dynaport Move Monitor and Actigraph, which are both devices that are worn on the hip. For further details on the study see (Van Remoortel et al. 2012).

The second study in WP2D was carried out as a follow up to the previous study. It was supposed to further assess the utility of activity monitors for use in clinical trials via a multicentre evaluation of the six commercially available monitors ('field study'). All tested monitors showed good correlations with 'active energy expenditure'. The best correlations were obtained with two of the triaxial monitors tested: the DynaPort MoveMonitor and the Actigraph GT3X. Another monitor, the 'Sensewear', (BodyMedia Inc) also passed all preset validation criteria. However this monitor is branded as a consumer device, rather than a medical device, and therefore was not further tested in subsequent PROactive-related studies. The DynaPort MoveMonitor and Actigraph GT3X monitors were also the best able to explain variability in total energy expenditure associated with PA, and were therefore most representative of what patients were actually doing. For further details on the study see (Rabinovitch et al. 2013).

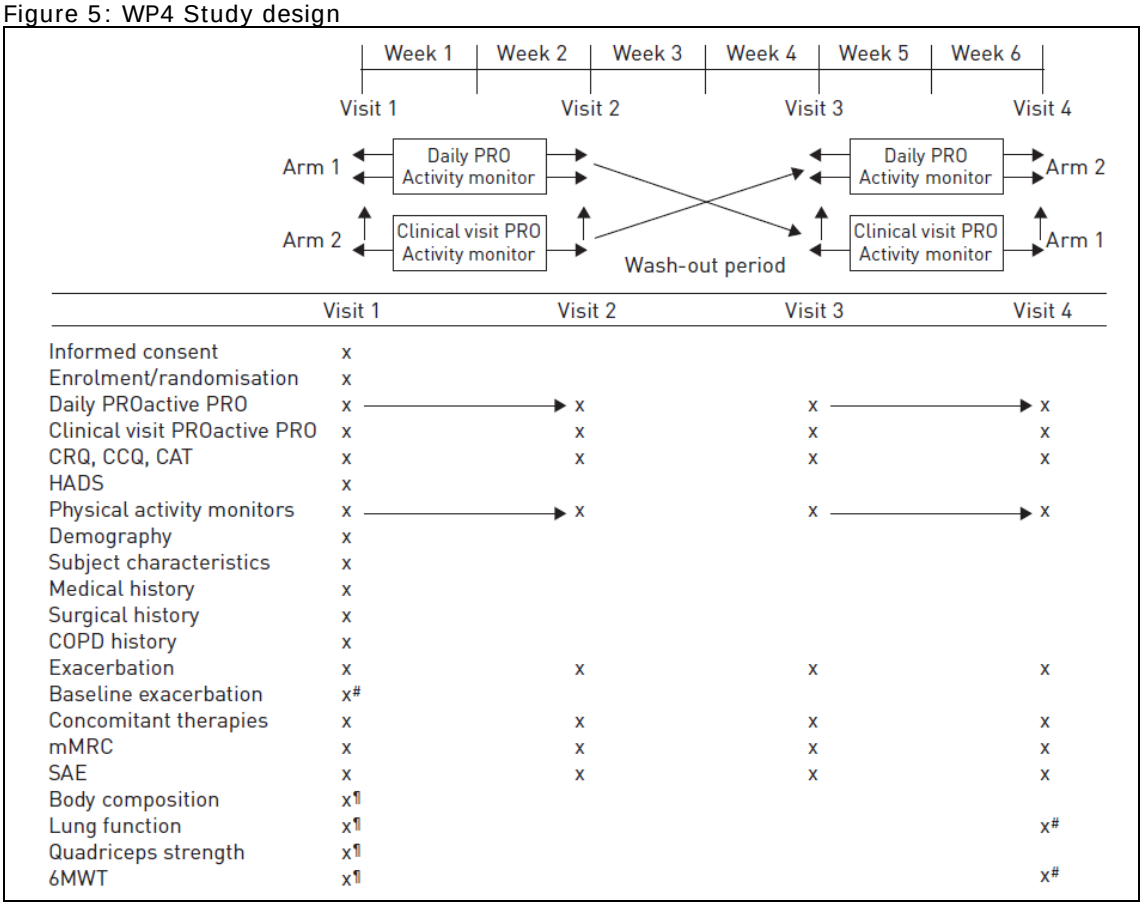
In summary, the data generated with these 2 studies, the laboratory validation study and the field study, have supported the use of the DynaPort MoveMonitor and the Actigraph GT3X in subsequent PROactive work packages WP4 and WP6 to further develop and validate the PROactive instruments.

Work package 4 (WP4) comprised an item reduction- and initial validation study with the primary objectives to

- derive the set of items that measure PA in both the daily and clinic visit versions of the PROactive instruments,
- confirm the draft PROactive conceptual framework of PA in patients with COPD for both the daily and clinic visit versions of the PROactive instruments,
- perform an initial validation of the two PROs instruments

The design of this multicentre study was randomised 6-weeks observation 2-way cross-over (Fig. 5).

Figure 5: WP4 Study design



	Visit 1	Visit 2	Visit 3	Visit 4
Informed consent	x			
Enrolment/randomisation	x			
Daily PROactive PRO	x	x	x	x
Clinical visit PROactive PRO	x	x	x	x
CRQ, CCQ, CAT	x	x	x	x
HADS	x			
Physical activity monitors	x	x	x	x
Demography	x			
Subject characteristics	x			
Medical history	x			
Surgical history	x			
COPD history	x			
Exacerbation	x	x	x	x
Baseline exacerbation	x#			
Concomitant therapies	x	x	x	x
mMRC	x	x	x	x
SAE	x	x	x	x
Body composition	x¶			
Lung function	x¶			x#
Quadriceps strength	x¶			
6MWT	x¶			x#

Both stable and exacerbated COPD patients were recruited, to cover the whole range of PA. In the first 2 week study period patients were randomised to complete either the daily PROactive item pool consisting of 30 questions asking patients to report their PA experience on a daily basis, or the clinical visit PROactive item pool of 35 questions using a 7 day recall. Following a 2 week wash-out patients completed the other questionnaire during the second study period. During the study periods, patients had to wear two accelerometers: the Actigraph G3TX and the Dynaport MoveMonitor.

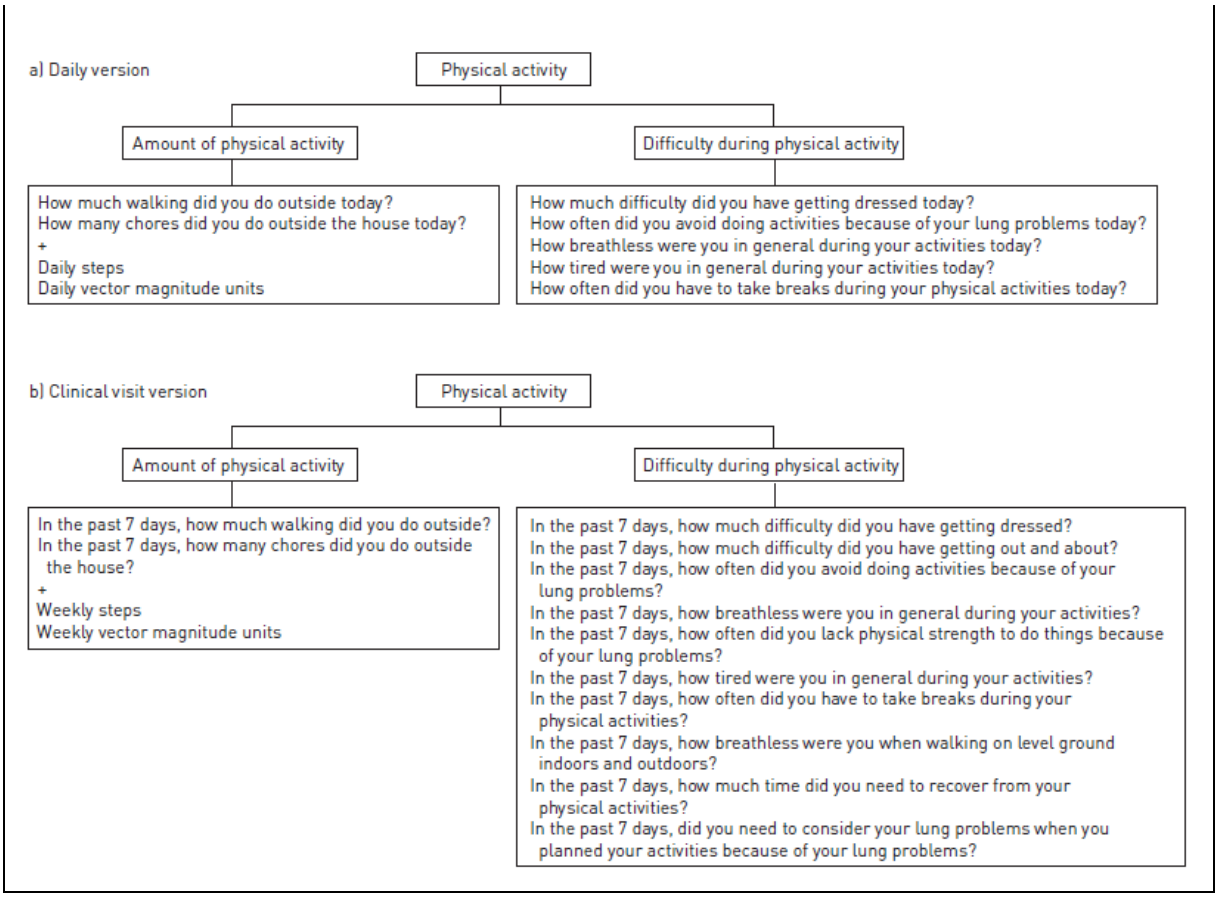
The design of the WP4 study was finalized following discussion with the QT which had some reservations regarding the adequacy of a cross-over design. However the Consortium's view was that the cross-over design allowed the use of a single large cohort, hence a broader range of COPD phenotypes to be included when compared with a two armed study using matched groups. The full cohort allowed for the evaluation of relationships between the two PROactive instruments of PA using paired data. Thirdly, the design lead to a substantial reduction in the burden of phenotyping these patients. The QT eventually agreed to the suggested design, also based on the review of draft versions of study protocol and statistical analysis plan (12, 13).

Two hundred and thirty six (n=236) patients with COPD were included in the WP4 study. Patients were mostly male (68%), with mean ±SD age of 67±8 years, FEV1 of 57±21% and body mass index of 27±5 $kg \cdot m^{-2}$. Most of them were GOLD II or III, 9% were GOLD IV, 46% had co-morbidity, and 60% had already been hospitalised for an exacerbation. A total of 228 patients (97%) had valid (≥3 days with ≥10 h wearing time) data from activity monitors, showing good compliance and moderate levels of PA.

For each of the two PROs two major methodological steps were carried out: domain identification was done first by exploratory factor analysis methods, which was then followed by domain-wise item reduction analyses (Rasch analyses). This sequential methodological approach actually carried out was sufficiently described and CHMP could finally support the Consortium's interpretation of the WP4 analyses' results. The analyses carried out suggested that both the daily and clinical visit versions of PPAC had a bi-dimensional structure, with a clear distribution of items in two factors. The two resulting domains 'amount of PA' and 'difficulties during PA' had been reported to be quite robust. As compared

to the preliminarily conceptual framework (figure 4), the revised conceptual framework (figure 6) no longer contains a symptom-specific domain, which indicates that the newly developed PROs have the potential to cover specifically the (isolated) concept of PA as targeted.

Figure 6: Conceptual frameworks of a) the daily version of PROactive Physical Activity in COPD (D-PPAC) and b) the clinical visit version of PROactive Physical Activity in COPD (C-PPAC) instruments: final domains and items



The resulting item sets as shown in the figure above were presented as 'draft PROs' after conduct of WP4. At the same time, the Consortium stated that no further changes in the PROs were foreseen at that point in time, and that all trials in WP6 were supposed to validate these very PRO versions. At that point in development the QT advised to maintain a certain amount of flexibility to amend/optimise the PROs (e.g. minor changes to response categories might turn out to be beneficial after broader use and testing). However, the Consortium stated that the items have been selected based on patient research and best statistical practice so should be robust going into WP6, where validation studies were planned to be running simultaneously, so timing of reporting would not permit adjustments of the PROs as part of WP6. For the QT, this fact constituted a minor deficiency in the PROs development and validation process. It was however understood that at least parts of the late phase validation trials would need to test and validate the final version of the PROs. As regards the intended implementation of monitor device data, the consortium considered different combinations of PRO question-items plus read-out variables from the activity monitors in the item reduction process. The two read-out variables 'daily steps' and 'mean Vector magnitude units per minute (VMU/min)' were found to be most informative in combination with the questionnaire items identified. Daily steps is understood to serve as a proxy for quantity of movement, whereas VMU serves as a proxy for overall intensity of effort. Cut-offs within the observed data ranges were chosen that maximised person separation index values in Rasch analyses. Interestingly, cut-offs differ between the two monitor devices investigated (Actigraph G3TX and the Dynaport MoveMonitor), which corresponds to a differential mapping from steps/day and VMU/min observed to PROs' response scores (0-4 or 0-5) finally assigned per monitor item included. Given that observation, it remained unclear for the QT in how far other monitoring devices than the two used in the validation trials could replace those monitors in the PROs without (repeated) thorough item-combination analyses including data cut-off investigations. The consequence is that the Opinion

given with this document is formally restricted to the PRO use involving either Actigraph G3TX and the Dynaport MoveMonitor. No recommendation is currently possible in relation to the use/implementation of other monitor devices in the data capturing of the D-PPAC and the C-PPAC.

Overall, CHMP agreed that the information presented indicate that a combination of monitor device read-outs and PRO items gives advantages in capturing amount of PA. Potential bias of wearing the monitor device on the actual amount of PA was discussed with the Consortium, and evidence exists that such bias might be negligible. The expectation that any potential bias of that kind would affect all parallel intervention (treatment) groups in a clinical trial in the same manner was acknowledged. Nonetheless, this general issue of biased estimation of PA might require dedicated consideration in the interpretation of future trial results.

Based on WP4 study data, some psychometric properties of the two PRO tools had been investigated.

According to the reports provided, both instruments showed strong internal consistency and test–retest reliability. Construct validity was explored via convergent-, known groups- and discriminant validity investigations. In both PROs instruments, the domain 'amount of PA' exhibited weak correlations with health-related quality of life and moderate correlations with dyspnoea and exercise capacity. The domain 'difficulty with PA', however, showed moderate to strong correlation with health-related quality of life, dyspnoea and exercise capacity. Known-groups validity was good in both instruments, with scores differentiating across grades of dyspnoea, stable from exacerbated patients at baseline and tertiles of PA levels (using variables not included in the PPAC scoring, such as intensity). Analyses for discriminant validity revealed low correlations with unrelated constructs.

For further details of analyses results see (14).

Throughout the qualification advice procedures, the question of whether the PROs should reveal one single total score each or, alternatively, separate scores for each of the two domains was repeatedly discussed. Based on the (early) descriptions of the Consortium's motivation to develop PROs to measure PA in COPD, the QT had a clear preference and advised to come up with one metric (per PRO) to describe PA as one entity. For the Consortium it was important to note that, according to their understanding, improving PA in COPD would either mean to improve the amount without negative impact on difficulty, or to improve difficulty without negative impact on amount, or to improve both amount and difficulty. With the advice provided, the QT saw no necessity to implement this 'restricted' definition of improved PA already into the scoring system of the PRO tools. It was felt that observed effects on a total score resulting from a mix of a slight negative change in one domain and substantial improvement in the other might still be relevant from a clinical perspective.

Such an understanding would be in line with the interpretation of the outcome of many other questionnaires (used in different disease areas) which feature more than one domain and one overall sum score. It is quite common that domain sub-scores are planned to be reported and interpreted in addition to allow for further exploration of the origin of observed effects. In the last round of discussion between the Consortium and the QT, the Consortium confirmed their concept to suggest the use of a total score (per PRO instrument), with the need to keep track of the two sub-domain scores. Both sub-domain scores are mapped to a range from 0 to 100 points, and the total score is derived by taking the average of the two domain scores. Hence, for each of the two PROs, the total score is also defined on the range from 0-100 points. Finally, agreement was reached that an overall effect in (perception of) PA may be driven by either or both domains, also reflecting the outcome of qualitative research with COPD patients.

With Work Package 6 (WP6) the PROs where further tested in clinical studies investigating the effect of different pharmacological and non-pharmacological interventions in patients with stable moderate to severe COPD, reflective of contemporary COPD management strategies (GOLD 2015).

With WP6, the Consortium was planning to address the following comments received in the final CHMP advice from the two qualification advice procedures:

- Interpretation of PRO results on PA has to be seen in the context of the pharmaceutical class of the drug used and the expected mechanism of action,
- Improved PA should not be at the expense of other aspects of Quality of Life (QOL) in COPD patients,
- The instrument may not be optimal for patients with milder COPD;

WP6 was therefore designed to:

- Confirm the internal consistency of the two PRO instruments
- Confirm test-retest reliability
- Evaluate and confirm construct validity
- Evaluate and confirm known groups validity
- Investigate the ability to detect change over time, i.e. the PROs' responsiveness
- Investigate these changes in relevant subgroups of patients, e.g. age, gender, COPD severity
- Determine the definition of response and investigate the minimal clinically important difference (MID)
- Verify the variables to use from the monitors and cut-offs from the activity monitors, and confirm the monitor outcomes as part of the PRO instrument scores.
- reconfirm the conceptual framework established after WP4

In line with WP6 objectives, the consortium has longitudinally validated the PROs in six clinical studies performed by EFPIA- and Academia partners. These studies are summarized below:

1. *PHYSACTO study*: An exploratory, 12 week, randomised, partially double-blinded, placebo-controlled, parallel group trial to explore the effects of once daily treatments of orally inhaled tiotropium + olodaterol fixed dose combination or tiotropium (both delivered by the Respimat® inhaler), supervised exercise training and behaviour modification on exercise capacity and PA in patients with COPD. The primary objective was to confirm that bronchodilator monotherapy (tiotropium) plus behavioural modification, bronchodilator combination therapy (tiotropium + olodaterol FDC) plus behavioural modification, and bronchodilator combination therapy (tiotropium + olodaterol FDC) plus exercise training plus behavioural modification improve exercise capacity as compared to placebo plus behavioural modification. The study population consisted of outpatients with COPD of either sex, aged 40 - 75 years with a smoking history > 10 pack years, post-bronchodilator FEV1 ≥ 30% and < 80% predicted, and post-bronchodilator FEV1/FVC < 70%.

2. *TRIGON - T9 study*: A Phase IIb, double blind, randomised, multinational, multi-centre, 2-way crossover, placebo controlled study designed to demonstrate the superiority of CHF 5259 (i.e. glycopyrronium bromide) vs. placebo, administered by pMDI over a 4-week treatment period in patients with moderate to very severe COPD (GOLD stage III and IV). Primary Outcome Measure was the change from baseline in pre-dose morning FEV1 on Day 28. Male and female adults (40 ≤ age ≤ 80 years) with a diagnosis of COPD being current or ex-smokers with a post-bronchodilator FEV1 < 60% of the predicted normal and a post-bronchodilator FEV1/FVC < 0.7 were included.

3. *URBAN TRAINING (CREAL) study*: This cross sectional and longitudinal RCT has – on top of validating the PROactive instrument - also provided opportunity to test an innovative intervention in patients with COPD. This study involved a training intervention adapted to each patient needs and capabilities and using public spaces and urban walkable trails. Primary objective was to assess 12 months effectiveness of the intervention with respect to PA level (primary outcome), and COPD admissions, exercise capacity, body composition, quality of life, and mental health (secondary outcomes) compared to "usual care". COPD patients aged >45 years with a ratio of forced expiratory volume in one second (FEV1) to forced vital capacity (FVC) ≤ 0.70 and clinically stable (i.e. least 4 weeks without antibiotics or oral corticosteroids) were included.

4. *ExOS study*: A cross-sectional and longitudinal open labeled 3 arm study was performed to primarily assess the functional capacity in patients with COPD and secure a wider understanding of the stability and sensitivity of commonly employed exercise tests so as to guide clinical trial outcome selection. This 7-9 week study compared the outcomes of the exercise tests following an (known) effective intervention, of either pulmonary rehabilitation or an inhaled bronchodilator (LAMA) therapy for 6 weeks. There was also a control arm with no intervention. Secondary objectives were to explore the relationship between PA and exercise testing and their responses to pulmonary rehabilitation and LAMA, and to report the MID of studied tests in response to pulmonary rehabilitation and LAMA. COPD patients with a GOLD stage 2-4 and MRC grade dyspnea 2 or greater, aged 40-85 years were included.

5. *MrPAPP study*: A cross sectional and longitudinal randomised clinical trial assessing the impact of a telecoaching program (COACH) on PA in patients with COPD on top of usual care, compared to usual care alone for 3 consecutive months. The COACH program included a step counter, an exercise booklet, an application installed on a Smartphone, the use of text messages and occasional telephone contacts with the investigator. PA was measured using the PROactive monitors (ActiGraph® and DynaPort®) and the PROactive questionnaire. A daily goal (number of steps) was

sent to the patient, and revised every week. Patients were 66 years old on average, with an FEV1=56±21% predicted, and 1/3 were female.

6. *ATHENS study*: Longitudinal randomised 4-arm study intended to compare paper-pencil versus the electronic scoring version of the PROactive instruments. All the patients who participated in the rehabilitation program were randomised in four groups: Group A included patients who only used the paper-pencil version of the clinical visit version of the PROactive instrument; in Group B patients used the electronic version of the clinical visit version of PROactive instrument; Groups C and D were used as control groups including patients who did not participate in a rehabilitation program while receiving the usual standard of care. Groups C and D were also randomized to those patients using the paper-pencil version (Group C) or the electronic version (Group D) of the PROactive instrument. The rehabilitation programme was multidisciplinary including mandatory supervised aerobic training 3 days a week, at appropriate training intensity, which was to be increased on a weekly basis. Resistance training was performed with fitness equipment also for 3 days/week. Other components of the program were breathing control and relaxation techniques, methods of clearance of pulmonary secretions, disease education, dietary advice, and psychological support on issues relating to chronic disability. Clinically stable patients with COPD were to be recruited from the academic centers' Outpatient Clinic on the following entry criteria if they had a post-bronchodilator FEV1 lower or equal to 70% predicted without significant reversibility (<12% change of the initial FEV1 value or <200 ml) and optimal medical therapy according to GOLD stage 2.

In the trials of WP6 D-PPAC and C-PPAC were implemented for use according to the descriptions as presented in Table 1.

Table 1: PPAC capture in WP6 individual trials

	PHYSACTO (BI)	**URBAN TRAINING (CREAL)**	**T9 TRIGON (Chiesi)**	**ExOS (UK NHS Trust)**	**Pulmonary Rehabilitation (ATHENS)**	**MrPAPP (Academic-TT)**
CT number	NCT02085161	NCT01897298	NCT02189577	-	NCT02437994	NCT02158065
N (included in analysis)	**283**	**308**	**161**	**33 (Pilot)**	**59**	**361**
Activity Monitor(s)	Dynaport	Dynaport	Dynaport	SenseWear & ActiGraph	Actigraph	Dynaport & Actigraph
Overall duration of study	19 weeks	12 months	12 weeks	7-9 weeks	8 weeks	3 months
PROactive	Key 2nd endpoint	Exploratory endpoint	Exploratory endpoint	Co-Primary endpoint	Primary endpoint	Key 2nd endpoint
D-PPAC	X	-	X	X	-	X
C-PPAC	-	X	-	-	X *	X
PPAC administration	•At Baseline, for 1 week (between V1 & 2) prior to randomisation at V4 •1st follow-up assessment: for one week between V5&6 in week 9 •2nd follow-up assessment: for one week between V7& 8 in week 12 PHT LogPad	At Baseline and At Month 12 Internet interface	Daily during 14 days during the run-in period for test-re-test purpose PHT LogPad	At Baseline and at the end of the study PHT Log¨Pad	At Baseline and at the end of the study Paper and computer version	1 week before randomization (V2) and at the end of study during week 12 (V3) PHT LogPad and Internet Interface

It should be noted that high-level data from two additional trials were expected to become available during the qualification procedure but have not been reflected on during preparation of this opinion document. (ACTIVATE Phase IV study evaluating a LABA/LAMA FDC (DUAKLIR®, GENUAIR®) in GOLD II-III COPD patients; AZ Phase IIa study in GOLD III-IV COPD patients with a history of frequent acute exacerbations with AZD7624, a new compound).

PHSYACTO, T9 TRIGON, EXOS and MRPAPP used/incorporated the D-PPAC. URBAN TRAINING, ATHENS and MRPAPP used/incorporated the C-PPAC. Both tools have accordingly been validated independently.

As has to be expected, adherence to protocol differed between trials and this resulted in only a part of patients contributing data to the final PROactive analyses for each trial (varying from 55% in study T9 TRIGON to 93% in PHYSACTO). Adherence criteria determining sufficient compliance for inclusion were set arbitrarily. For validation purposes it is endorsed to focus on a sample indeed contributing data points. No comparison of baseline characteristics between adherers and non-adherers were performed and the possibility of systematic exclusion of certain patient groups (e.g. based on severity of impairment) from the validation exercise cannot be fully ruled out. At the same time, it is understood that the baseline and EOT data reported only reflect those patients eventually included in the analyses which mitigates respective concerns.

Key demographics were largely comparable across trials and agreeably representative of a COPD population. Overall, about half of patients were younger than 65 years, about two thirds were male. Participants were predominantly non-smoking, retired and not living alone.

Table 2: Baseline demographics and comorbidities in WP6 trials

	Total	Physacto	T9	Exos	MrPapp	Urban Training	Athens
	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)
n	995 (100)	195 (20)	88 (9)	22 (2)	330 (33)	308 (31)	52 (5)
Age (years)	66 (8)	65 (6)	62 (8)	64 (7)	66 (8)	69 (9)	67 (9)
Age groups:							
<65	384 (39)	78 (40)	55 (63)	10 (45)	131 (40)	89 (29)	21 (40)
≥65 & <75	445 (45)	114 (58)	27 (31)	11 (50)	145 (44)	131 (43)	17 (33)
≥75 & <85	159 (16)	3 (2)	6 (7)	1 (5)	53 (16)	82 (27)	14 (27)
≥85	7 (1)	0	0	0	1 (0.3)	6 (2)	0
Male	701 (70)	124 (64)	52 (59)	17 (77)	209 (63)	257 (83)	42 (81)
Socioeconomic status: low	256 (67)			11 (50)		208 (68)	37 (71)
Living alone	167 (23)			5 (23)	83 (25)	40 (13)	39 (75)
Active worker	90 (13)			3 (14)	45 (14)	36 (12)	6 (12)
Current smoker	302 (30)	76 (39)	57 (65)	3 (14)	85 (26)	71 (23)	10 (29)
Weigth (kg)	77 (17)	79.9 (17.7)	78.6 (18.5)	75.2 (20.4)	75 (16.7)	77 (14.9)	77.6 (17.5)
Heigth (cm)	168 (9)	170 (10)	170 (9)	169 (10)	168 (9)	164 (7)	168 (9)
BMI (kg/m^2)	27.3 (5.1)	27.4 (4.8)	27 (6)	26 (5.7)	26.4 (5)	28.4 (5)	27.3 (5.1)
Heart Rate (bpm)	76 (13)	73 (11)	72 (11)	79 (11)	77 (13)	77 (14)	78 (11)
Systolic BP (mmHg)	136 (18)	133 (18)	132 (10)	125 (20)		140 (18)	
Diastolic BP (mmHg)	79 (11)	78 (10)	80 (8)	80 (14)		79 (11)	
Doctor diagnosed co-morbidities							
Anxiety	10 (2)	7 (4)	0	0	3 (1)		
Depression	44 (7)	27 (14)	0	3 (14)	14 (4)		
Cancer	26 (4)	14 (7)	0	2 (9)	10 (3)		
Any cardiovascular	149 (23)	66 (34)	18 (20)	3 (14)	62 (19)		
Diabetes	45 (7)	13 (7)	0	5 (23)	27 (8)		
Musculoskeletal	206 (20)	89 (46)	3 (3)	3 (14)	65 (20)		
Asthma	3 (1)	0	0	1 (5)	2 (1)		
Hypertension	183 (29)	92 (47)	0	12 (55)	79 (24)		

Some variables have missing data: 614 in Socioeconomic status, 283 in Living alone, 292 in Active worker, 464 in co-morbidities, 4 in HR, 385 in BP.

Relevant co-morbidities are listed in Table 2 as well. Importantly, keeping in mind the patient demographics, concomitant musculoskeletal disorders seem underrepresented in some of the trials, or respective data are missing (UT, Athens trials). Drawing on the inclusion/exclusion criteria of the concerned trials, all but one trial (i.e. T9 Trigon) explicitly exclude concomitant conditions that could interfere with PA, including orthopaedic, neurological but also, more generally, “other” respective complaints unrelated to COPD. Whereas it is evident that concomitant diagnoses interfering with a patients activity level would hamper demonstrating PPAC performance related to pulmonary activity limitations or improvement thereof, this might have created a somewhat artificial setting. As seen in the table above, the exclusion criteria did not prevent all patients suffering from potentially relevant conditions from entering the trials. Still, whether the PPAC tools would perform similarly (well) in a broad COPD population without abovementioned restrictions as regards co-morbidities in terms of staging COPD-related PA and being responsive to pulmonary improvement cannot conclusively be answered.

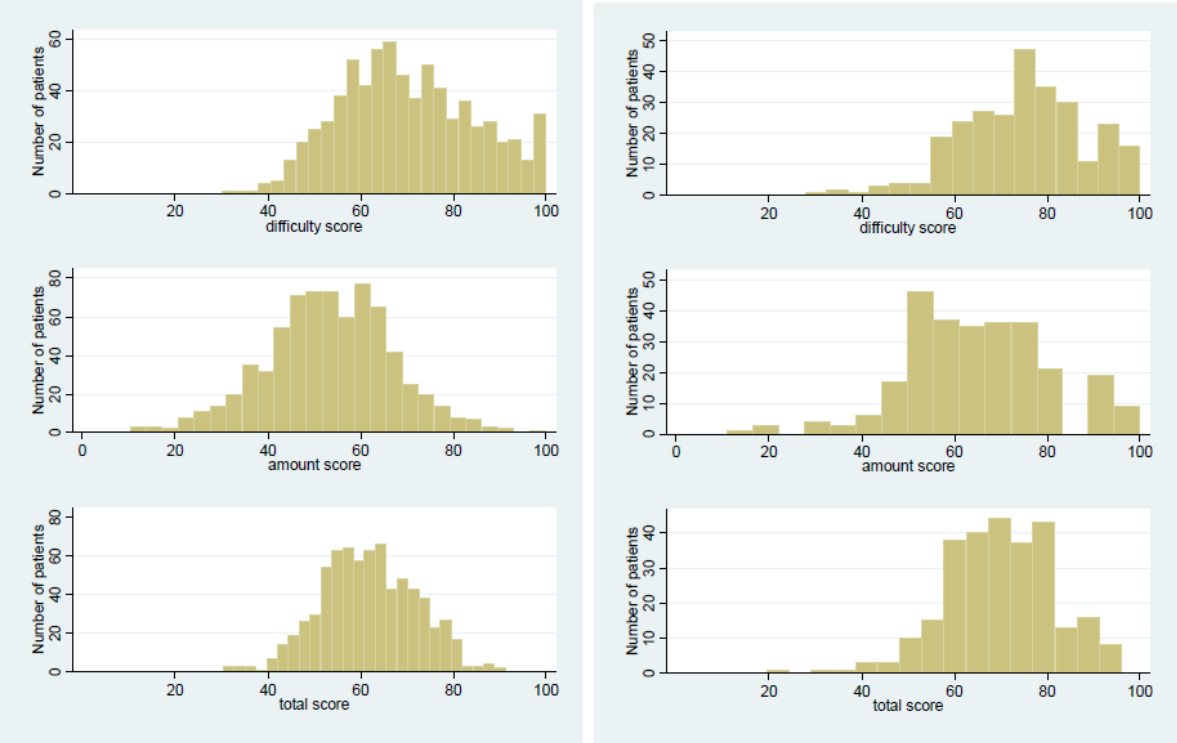
Table 3: Baseline COPD/physical activity in WP6 trials

	Total	Physacto	T9	Exos	MrPapp	Urban Training	Athens
	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)	m (SD) / n (%)
n	1083 (100)	283 (26)	88 (8)	22 (2)	330 (30)	308 (28)	52 (5)
Lung function							
FEV_1 (% predicted)	53.1 (18.1)	48.3 (12.7)	48.3 (12.1)	46.2 (19.6)	56.6 (21.7)	55.8 (17.79	51.3 (19.8)
ATS/ERS stages:							
Mild: $FEV_1 \geq 80\%$	86 (8)	1 (0.4)	1 (1)	0	53 (16)	27 (9)	4 (8)
Moderate: $FEV_1 < 80\%$ & $FEV_1 \geq 50\%$	493 (46)	124 (44)	41 (47)	8 (36)	136 (41)	161 (52)	23 44)
Severe: $FEV_1 < 50$ & $FEV_1 \geq 30$	409 (38)	139 (49)	41 (47)	9 (41)	104 (32)	95 (31)	21 (40)
Very Severe: $FEV_1 < 30$	95 (9)	19 (7)	5 (6)	5 (23)	37 (11)	25 (8)	4 (8)
FVC (% predicted)	88.7 (22.7)	104.1 (20.4)	79.7 (14.5)	82.6 (13.6)	92 (23.2)	75.6 (17)	79.4 (18.9)
FEV_1/FVC (%)	49.7 (13.4)	47 (10)	48 (13)	43 (14)	49 (15)	54 (12)	48 (14)
IC (% predicted)	76 (29)	72 (17)	68 (15)		86 (40)		73 (33)
SaO_2 (%)	95 (2)	96 (2)		96 (2)	95 (2)	95 (2)	95 (2)
Exercise capacity & muscle strength							
6MWD (m)	456 (103)	452 (100)		424 (81)	443 (105)	485 (94)	400 (113)
ESWT time (s)	303 (207)	295 (199)		401 (287)			
Patient Global Rating PA							
Not limited	169 (26)	50 (18)			87 (27)	32 (62)	
A little bit limited	324 (49)	162 (58)			146 (45)	16 (31)	
Limited	142 (22)	59 (21)			80 (25)	3 (6)	
Severely limited	21 (3)	9 (3)			11 (3)	1 (2)	

With regards to baseline lung function and exercise capacity the large majority of patients can be classified as GOLD 2/3, showing some degree of limitation regarding PA. Whereas this is expected to represent the COPD population at large, it is noted in the context of validating an outcome tool that for lung function patients at both ends of the scale are not well represented and for PA this particularly applies for those being severely limited. The 6MWD averages also indicate a reduced, yet considerable residual performance level. Accordingly, the Applicant stated that at the current stage, very severe COPD and/or patients currently suffering from an exacerbation (implying a rather dynamic disease state) are not considered a target population for applying the PPAC outcome tools.

Two (likely interdependent) observations can be made regarding the distribution of baseline D-PPAC and C-PPAC scores in the aggregated study sample:

Figure 7: Distribution of D-PPAC scores (left panel) and C-PPAC scores (right panel) at baseline



Firstly, it appears that, in line with statements made above on the disease severity of included subjects, no patients scored at the lower end of either D-PPAC or C-PPAC in any of the clinical studies. This applies to all three scales ('difficulty', 'amount' and 'total'), but is most pronounced for the 'difficulty' and 'total' scales where apparently no subjects scored below 40 (out of 100) and the majority substantially higher. This means that the psychometric properties of the tools at the lower end of possible scores were essentially left unaddressed during the WP6 validation exercise. Secondly, when looking at known-groups validity, i.e. comparing PPAC scores with GOLD stage at baseline, it appears that while showing variably pronounced separation in PPAC scores depending on GOLD stage, even those patients with substantially impaired lung function (i.e. GOLD 4) scored relatively well on D-PPAC and C-PPAC. The same holds true for dyspnoea (mMRC) and 6MWD results if employed as well-known group denominators. Whereas these observations might be explained by patient selection, populations appear rather comparable between the WP4 study conducted for initial validation and item reduction and the WP6 trials, and the existence of a floor effect cannot be ruled out.

Given the differences in trial designs and PPAC data capture schedules, the different trials were not equally able to contribute information for all the validation sub-tasks as listed in Table 5 below. From the different trials, PRO response data of similar structure were pooled to obtain new datasets, each one eventually foreseen for a specific part of the validation analyses.

For the D-PPAC three different datasets were derived for different validation analysis tasks:

- 'PDDR'-dataset: Pooled Daily PPAC day-by-day retest, to analyse Test-retest reliability;
- 'PDRB'-dataset: Pooled Daily PPAC Random baseline, to test Construct validity and confirm the conceptual framework;
- 'PDRR'-dataset: Pooled daily PPAC Random repeated, to analyse responsiveness;
- For the C-PPAC two different datasets were derived for different validation analysis tasks:
  - 'PCB'-dataset: Pooled Clinical visit PPAC Baseline, to analyse Internal Consistency, Construct Validity, and to confirm the conceptual framework;
- 'PCR'-dataset: Pooled Clinical Visit PPAC Repeated, to analyse responsiveness;

Details on the data-pooling/data-merging approaches are provided in the Statistical Analysis Plan of WP6 (15). The data management in this context was adequately described and documented, and the data-sets used as basis for different validation tasks were considered suitable by the QT.

One further important aspect in relation to the handling of data captured by the D-PPAC and C-PPAC is the standardised approach of actual data aggregation. It was agreed with the Consortium that qualification can only be considered for the format of data aggregation used in development and validation of the tools.

For the D-PPAC the intention is to derive weekly averages, based on daily recordings and the need to merge on a daily basis:

- Response to valid daily questionnaire (no missing answers)
- Values of steps and VMU/min if valid activity monitor data (valid means at least 8h of monitoring)
- calculate daily amount, difficulty and total score
- calculate weekly mean if at least for 3 days in the week the questionnaire and monitor data are available; data from days were only questionnaire data or only monitor data are available are not taken into consideration for calculation of scores;

For the C-PPAC the intention is to use one weekly single measure, based on

- Response to valid clinical visit questionnaire (no missing answers)
- Median values of steps and VMU/min of three to seven valid days prior to clinical visit questionnaire (at least 8h per monitoring day irrespective of weekdays/weekends)
- Calculation of amount, difficulty and total score

One finding in the review of WP6 data was the rather divergent estimation of 'baseline' data in the MrPAPP trial, dependent on which PPAC tool was used for data capture. The MrPAPP trial was the only WP6 study in which both PROs were scored at baseline. According to the study results provided, the PROs score 8-10 points differently on average in the same study population. Although the actual patient set used was not identical for the two PROs to derive total scores (different 'n' obviously due to differences in missing data structure), the differences seen in average scores are quite extensive, so that interchangeable use of the two PROs within one trial setting cannot be supported based on these findings.

Patient compliance to the PROs was another topic discussed in the framework of the qualification procedure. Given the hybrid nature of the two tools (monitor + questionnaire data required from the same data capture period/days), there is in principal an elevated risk for lower patient compliance if data capturing is relying on more than one source. However, the Consortium concluded from the different WP6 trials that in general compliance increased with 'importance' of measuring PA in the specific trial setting. In this context, it is important to note that, whenever one of the two PRO tools is intended to be applied, investigators and study personnel need to be adequately trained to use/introduce the PPAC in a specific trial. This is expected to positively impact patient compliance. Of course, also on the patient side, there is a need to provide appropriate information on how PPAC – related activities are supposed to be handled during the conduct of the trial. For all of these purposes, the adequacy of the User's guide (16) is of importance.

Data capture for the D-PPAC is supposed to be done with an electronic hand-held device. Relevant experience was gathered in the clinical WP4 and WP6 trials. Questions regarding device selections as well as questions relating to technical validity/performance where not directly addressed in the framework of the qualification procedure. For the C-PPAC, a paper and pencil version as well as a web-based interface was developed and tested by the Consortium. As for the D-PPAC, technical details to support the electronic version of the C-PPAC have not been subject to assessment in this qualification procedure.

So far, the D-PPAC in available in 62 languages whereas the C-PPAC is available in 14 languages. Translation programmes included cognitive interviews performed with patients having the corresponding language as mother tongue. Assessment of the translation work was not subject to this qualification procedure.

Reliability, construct validity and responsiveness of both D-PPAC and C-PPAC were investigated in WP6 as outlined below:

Table 4: Psychometric properties tested per study

		EFPIA			Academic		External
		Physacto	T9	EXOS	MrPaPP	Athens	UT
PROactive instrument	Daily	X	X	X	X		
	Clinical visit				X	X	X
	Dynaport	X	X		X	X	X
	Actigraph			X	X		
Reliability							
Internal Consistency		X	X	X	X	X	X
Test-Retest Reliability			X				
Construct validity							
Convergent validity		X	X	X	X	X	X
Discriminant validity		X	X	X	X	X	X
Known groups validity		X	X	X	X	X	X
Ability to detect change (responsiveness)		X		X	X	X	
Confirmation of Conceptual Framework		X	X	X	X	X	X

Psychometric properties D-PPAC:

As regards reliability measures, internal consistency and test-retest reliability were addressed. Crohnbach's alpha was consistently >0.7 for both 'difficulty' and 'amount' domains in the total dataset and in each of the 4 included studies. Test-retest reliability was tested using Intraclass Correlation Coefficient values and Bland Altman plots. Only data from the T9 TRIGON study were used since it was the only study that had repeated measures within a range of 7 (+/-1) days. Analysis was done by comparing average measures of Week 1 with those of Week 2 but it should be noted that patients were subjected to a change in medication at the beginning of week 1 compared to baseline. Results (suggesting a high correlation) therefore have to be interpreted with caution, also because this strategy was apparently chosen over a comparison of day 6 vs. day 13 scores in a data-driven manner.

Construct validity was addressed via correlation with related and unrelated constructs and with known-groups expected to have differences in PA. Convergent validity was tested against different known measures of dyspnoea, health status, exercise capacity and PA. Correlations were modest and varied widely depending on domain and related construct applied and the Applicant attributes this to the fact that PROactive instruments measure different concepts than already existing instruments, which is difficult to ascertain. It is noted that for 'global rating of PA', a presumably simple construct, good correlation with the PROactive instruments across domains would have been expected which was apparently not the case. Expectedly unrelated constructs (i.e. height, heart rate, BP) were found to not correlate with PPAC scores. As already stated above, known groups comparisons support the differentiation of impairment severity via D-PPAC but only so over a limited range of the scale.

Caution is warranted regarding interpretation of responsiveness because clinical trials included in this analysis did not include interventions of known efficacy. Thus, the PRO may falsely seem not responsive, when the interventions are not effective. According to the Applicant, EXOS study results were removed from responsiveness analysis because only 22 patients (distributed in 3 different groups) participated. In PHYSACTO the response was more pronounced across all three domains in all interventional arms tested, compared to the placebo arm. In MrPapp no change from baseline was observed for either arm with the 'amount' domain being the sole exception where minor improvement was observable for the telecoaching intervention and minor worsening for the usual care arm.

For the investigation of longitudinal validity, MrPaPP and PHYSACTO data were pooled and three variables of self-reported global rating of change were categorised and possible responses to each were grouped as follows:

- Global rating of change 'difficulty'
  - much more difficult, more difficult, a little more difficult
  - no change, a little easier
  - more easy, much more easy
- Global rating of change 'amount'

  - much less active, less active, a little less active
  - no change, slightly better
  - more active, much more active
- Global rating of change 'overall'
  - much worse, worse, slightly worse
  - no change, slightly better
  - better, much better

Whereas the grouping of response possibilities into -/=/+ can be criticised as it limits a further differentiation for quantity of change, the direction of effect as evident from all three D-PPAC domains was concordant for each category of global rating.

Furthermore, differences between final and baseline PA levels were calculated using variables from the activity monitors not included in the calculation of PPAC scores, including time in light, moderate and vigorous PA, intensity, and lying, sitting, standing and walking time. According to the distribution of the differences and their values, the following variables were used for longitudinal validity: changes in time in moderate-to-vigorous activity, changes in time lying or sitting and changes in intensity and each categorised in quintiles. Only results on change in time in PA are provided which support the assumption of the scales being responsive, at least for the 1st and 5th quintiles, i.e., in those patients with most increase or reduction in time in PA. Finally, 6MWD changes were compared to PPAC changes and results indicate that concordance in response was only there for those patients increasing their walking distance but not for those showing a reduction as in these patients PPAC scores stayed stable over time.

For determining a potential MID of the D-PPAC, anchor-based as well as distribution-based methods were used relying on PHYSACTO and MrPapp data. 6MWD, CCQ and SGRQ as well as change in global rating ('total', 'difficulty' and 'amount') were considered as established outcomes that could serve as candidate anchors. Correlations between these candidate anchors and the three PPAC domains were however rather low, somewhat surprisingly also so for change in global rating. Since there are three categories of GRCs (worse, no change or little easier, better), the mean change in the amount score in patients which reported an improvement in the global ratings of change was chosen to represent the MID. In order to be consistent with the estimation of MIDs based on GRC the mean change in the difficulty score in patients who had improvement in the CCQ of at least -0.4 (MID of CCQ (Kocks et al. 2006) or of at least -4 (MID of SGRQ - Schünemann et al. 2003) were selected as MIDs. For the GRC the mean change in the difficulty score in patients who reported an improvement in the GRC difficulty was considered to represent the MID. 6MWD was disregarded for the low correlation with PPAC scores. The obtained MID estimates were between 5.2 and 7.8 for the difficulty score and 4.7 and 6.7 for the amount score. The anchor- and distribution based methods yielded similar results but it is noted that SDs were quite large. Based on that, a MID of 6 for the amount score and a MID of 6 or 7 for the difficulty score was deemed optimal. In order to simplify the interpretation it was suggested to use a MID of 6 for both scores of the D-PPAC. For the total score the MID estimates were between 2.0 and 5.7. For this score it was suggested to use a MID of 4.

The anchors and their respective MIDs used seem reasonable based on the cited literature but the low correlations with PPAC and the assumed independency of concepts clearly renders "global rating" anchors more meaningful than others. Derived estimates for MIDs for 'amount' and 'difficulty' derived showed some differences and where pragmatically and uniformly set across tools and scores for reasons of simplification. In this context it is noted that the Company states: "PA can be considered relevant (i) when a given improvement in amount is achieved without more difficulty, (ii) when less difficulty with PA occurs without deterioration in the amount, or (iii) both less difficulty with activity and a greater amount of activity are demonstrated." This simple approach can be followed to jointly consider the 'amount' and 'difficulty' domains in specific scenarios but does not consider situations where certain deteriorations in either domain might be accompanied by substantial gains in the other (which could result in a net benefit). The 'total' domain combining amount and difficulty can be a remedy but the lower proposed MID is clearly questioned as less than meaningful improvement on either amount or difficulty paired with no change in the respective other domain, could be considered meaningful in the total scale which is counterintuitive. Overall, how certain changes in the three domains would be perceived by the patient, likely also depending on baseline values, seems not conclusively answered. MID determination usually focuses on the immediate benefit associated with certain quantitative changes in the concerned score rather than the predictive value of such changes for other (preferably long-term) outcomes with established or intrinsic clinical relevance such as survival. The latter however also constitutes a viable strategy for making PRO outcomes interpretable

and informative for benefit assessment of experimental interventions. So far, the predictive properties of certain baseline and/or changes in PPAC or subdomains for e.g. survival, dependency, or lung outcomes such as exacerbations, etc. were not investigated during validation. Feasibility constraints for such analyses are acknowledged however, at least for survival, looking at the duration/size of studies included in WP6.

Psychometric properties C-PPAC:
As regards reliability measures, only internal consistency was addressed on MrPapp and UT data. Test-retest reliability was not studied because of the design of the included studies. None of the studies included a repeated questionnaire one week apart. Crohnbach's alpha was consistently >0.7 for both 'difficulty' and 'amount' domains in the total dataset and in each of the 2 included studies.

Construct validity was addressed via correlation with related and unrelated constructs and with known-groups expected to have differences in PA. Convergent validity was tested against different known measures of dyspnoea, health status, exercise capacity and PA. As seen for the D-PPAC, correlations were modest and varied widely depending on domain and related construct applied and the Applicant attributes this to the fact that PROactive instruments measure different concepts than already existing instruments, which is difficult to ascertain. It is noted that for 'global rating of PA', a presumably simple construct, good correlation with the PROactive instruments across domains would have been expected which was apparently not the case. Expectedly unrelated constructs (i.e. height, heart rate, BP) were found to not correlate with PPAC scores. As already stated above, known groups comparisons support the differentiation of impairment severity via C-PPAC but only so over a limited range of the scale.

MrPapp and ATHENS data were used to analyse responsiveness of the C-PPAC. The Athens study was a 4-arm study designed to compare the paper-pencil with the electronic version of the PROactive instrument. With this study, the patients who participated in the rehabilitation program were randomized in four groups: Group A included patients who only used the paper-pencil version of the clinical visit PPAC; in Group B patients used the electronic version of the clinical visit PPAC; Groups C and D were used as control groups with patients only receiving the usual standard of care. In both trials, the intervention arms displayed higher response across C-PPAC domains compared to control. The control arms, particularly those in the ATHENS trial, also reflected varying degrees of worsening across domains. Overall, and as seen for the D-PPAC, C-PPAC seems capable of reflecting changes to PA. The subjects dealing with the paper version showed more marked response (in both directions) than those dealing with the electronic version, but no formal comparison of the two modalities was made.

For the investigation of longitudinal validity, only MrPaPP data are referred to and three variables of self-reported global rating of change were categorised and possible responses to each were grouped in the same manner as described above for the D-PPAC. The direction of effect in all three C-PPAC domains was concordant with each category of global rating, thus supporting the notion of longitudinal validity. Furthermore, as for the D-PPAC, differences between final and baseline PA level and 6MWD were calculated and grouped in quintiles. Concordance with C-PPAC changes can be observed with exception of the 'difficulty' domain not reflecting changes in PA which can however potentially be explained by patients adapting 'amount' while maintaining stable levels of 'difficulty'.

For MID determination, same methods as for the D-PPAC were used but only MrPapp data were considered. As seen for the D-PPAC, correlations between candidate anchors and the three PPAC domains were rather low. The MID estimates ranged between 2.8 and 6.8 for the difficulty score and 4.5 and 7.9 for the amount score across anchors, all estimates with little precision. A MID of 5-6 was considered appropriate for the amount and difficulty scores of C-PPAC by the Applicant, but 6 was kept for reasons of consistency with the D-PPAC. For the total score the MID estimates ranged between 3.4 and 5.9. For this score it was also suggested to use the same MID of 4 as for the D-PPAC. The critical discussion provided above on MID derivation applies similarly for the C-PPAC.

For further details of analyses results of WP6 see (17).

During the Qualification procedure the topic of the 'Context of Use' for the two different PRO tools (separately) was further discussed with the Consortium. The idea was that the choice of Daily or Clinical Visit tool is driven by the clinical hypothesis being tested and therefore the study design. The suggestion for the C-PPAC was that it would more likely be used where patients' experience of PA is a supportive outcome and/or where patient burden of completing PROs is high. In relation to the

intended use of the C-PPAC, also 'pragmatic studies' to gather real-world evidence (e.g. 'minimal' intervention studies) where suggested. The D-PPAC was suggested to be used in the context of study settings where measurement of patient experience of PA is an outcome of primary interest. The Consortium's idea was that whenever "label-claims" could result from PA data analyses, the basis for calculation should be the D-PPAC.

Finally, the presented results of WP6 could be continuously updated/confirmed with data from trials still ongoing during consultation or planned for the future. It is further stated that stratified results for all validity analyses are available (i.e. based on gender and COPD staging) and respective high-level data might also be useful for public domain to support applicability of PPAC across relevant substrata.

CHMP opinion

The Consortium developed two PRO tools, the D-PPAC and the C-PPAC to capture physical activity (PA) data in patients with COPD in clinical trial settings. Both tools are hybrid tools, combining information from questionnaire items with PA monitors read-out data. State-of-the-art qualitative methodology has been applied in the development phase to build a conceptual framework that eventually combines two domains: 'amount of PA' and 'difficulty with PA' into one concept for each of the two PRO tools. This resulting conceptual framework is seen reasonable to describe PA in COPD. In general, also adequate quantitative methods have been used to identify the optimal sets of items, monitor read-outs and response categories which finally comprise the D-PPAC and the C-PPAC.

With a recall period of 24 hours the D-PPAC allows to collect data on a daily basis. Derived data is converted to two domain scores and one total score, covering average information for one week. The recall period as well as the actual data aggregation approach is endorsed for the use of the D-PPAC. It is agreed to the consortium that due to the higher amount of information collected on a daily basis, the D-PPAC qualifies for a context of use where a clear (primary) focus is on measuring PA. In the decision to apply the D-PPAC in a specific study, expected patient burden should however be considered and weighed against the importance of PA as study objective.

The C-PPAC has a recall period of 7 days, which is indeed considered an adequate period to capture PA data reflecting weekly (repeated) routines of COPD patients' daily life. As for the D-PPAC, data is converted to two domain scores and one total score for the C-PPAC. The suggested context of use for trial settings where patients' experience of PA is a supportive outcome and/or where patient burden of completing PROs can be expected to be high can be endorsed in principle.

It is important to note that for both tools, the D-PPAC and the C-PPAC, the derived *total* score is the simple average of the two domain scores (amount & difficulty), giving the two domains equal weights in computation. This is considered a simplistic approach, and there is currently no evidence from the validation work that the equal-weights approach is most reasonable to derive a total score to describe PA as desired. As during development item selection/optimization was done separately for the two domains, there is no overall (items) evaluation of optimality regarding the PROs' single components (i.e. items and monitor read-out data). Against this background, it seems advisable to focus eventual interpretation of PRO results on the two resulting domain scores for 'amount' and 'difficulty' next to each other, rather than on the *total* score. Further development work seems indicated to pursue the goal of having a total score being most informative for PA in the trial settings targeted.

The Consortium's validation work contains an attempt to determine minimal important differences (MIDs) on the PROs' domain- and total scales. Whilst anchor-variables and their respective MIDs seem reasonably selected, low correlations between some of the anchors and PPAC scores were observed. These finding might just reflect the fact that PA - as the new entity of interest - is indeed rather independent from other established measures commonly used in COPD. Derived estimates for MIDs for 'amount' and 'difficulty' were pragmatically and uniformly set across tools and scores for reasons of simplification. Uncertainty remains how certain changes in the PRO domains would be perceived by the patient (likely also depending on baseline values). MID determination focused on the immediate benefit associated with certain quantitative changes in the concerned score rather than the predictive value of such changes for other (preferably long-term) outcomes with established or intrinsic clinical relevance (e.g. survival). The latter however also constitutes a viable strategy for making PRO outcomes interpretable and informative for benefit assessment of experimental interventions. So far, the predictive properties of certain baseline and/or changes in total PPAC or domains for e.g. survival or lung outcomes such as exacerbations etc., were not investigated.

In the validation work for the two tools, psychometric properties were evaluated on basis of patient sets which excluded individuals who might have scored at the lower end of the domain/total scales. This means that interpretation of derived psychometric properties for the two tools is limited to data ranges corresponding to central and upper parts of the underlying score-data distributions. In how far this corresponds to restrictions in targeted COPD patient population or is related to a potential floor effect of the tool (i.e. being insensitive to differentiate among worse PA scores) remains currently unclear (e.g. GOLD 4-categorised patients were found to score relatively high on the D-PPAC and C-PPAC). Patients with relevant comorbidities potentially interfering with PA have also been systematically excluded from validation trials which might either require further restrictions or careful interpretation of PA data collected in such patients.

Data from validation trials suggest that the D-PPAC and the C-PPAC cannot be used interchangeably in one trial, as averaged domain/total scores on the group level might not be in the same range based on the same time-period for data capture.

From a technical perspective, the Opinion provided here is formally restricted to the PROs' use involving either Actigraph G3TX and the Dynaport MoveMonitor. No recommendation is currently possible in relation to the use/implementation of other monitor devices in the data capturing of the D-PPAC and the C-PPAC.

The original Consortium's request for Qualification opinion contained a suggestion for two Clinical trial endpoint models where the new C-PPAC and D-PPAC were proposed to be used as secondary, or even as primary efficacy endpoints in COPD trials. The current EMA *Guideline on clinical investigation of medicinal products in the treatment of COPD* (EMA/CHMP/483572/2012-corr) mentions PA as a potential secondary endpoint, and contains clear recommendations regarding primary endpoints to be envisioned in various study/patient population settings. During the qualification review it became clear that discussion around clinical endpoint models and potential positioning of PA in the hierarchy of important endpoints in COPD trials should be kept separate from the actual qualification aim to declare the two new PRO tools principally suitable to capture PA in COPD patients as intended. It was decided to strive for qualification without touching the issue of whether the PROs are suitable to inform (co-) primary/secondary (etc.) endpoints in the various suggested contexts of use. Against this background it is important to note that any descriptions contained in the User Guide which refer to positioning of endpoints and targeted claims have not been discussed/agreed in the margins of this qualification procedure.

Incorporating findings based on the PRO tools in 5.1 of the SPC of a compound targeting COPD seems possible but specific content or wording cannot be pre-empted at this point in time and will largely depend on the effects shown in a specific development programme and the perceived relevance of such information to the patient/prescriber, accounting for overall results. As discussed above, the interpretation of certain changes observable on PPAC and its subdomains in terms of magnitude and associated patient-perceived benefit is considered difficult and might require further context, i.e. embedding in other (secondary) outcomes.

In the framework of the qualification advice/opinion procedures, there was no dedicated assessment of technical details of electronic formats for the D-PPAC (hand-held) and the C-PPAC (web-based solution). It is also important to note that translation work carried out for the two PRO tools was also not subject to this qualification procedure.