Guideline on Using Patient Reported Outcomes in Drug Clinical trail

(Trial)

Center for Drug Evaluation, NMPA

December, 2021

Table of Contents

Guideline on Using Patient Reported Outcomes in Drug Clinical trail

I. INTRODUCTION

Clinical outcome is the core basis for evaluating the benefits and risks of drug treatment, how to observe clinical outcomes accurately, reliably, and completely is critical. Patient-reported outcome (PRO) is one of the forms of clinical outcome. It has been used more and more widely in clinical research of drug registration. In addition, as the concepts and practices of Patient-Focused Drug Development (PFDD) continue to evolve, there is an increasing focus on acquiring data about patients' experiences, opinions, needs during the entire life cycle of drugs and effectively integrating them into drug development and evaluation. Clinical outcome assessments (COA), especially PRO, can reflect the feelings of patients. It is an important part of PFDD.

This guideline aims to provide sponsors with guiding opinions for the rational use of PRO data in drug registration research from the following aspects, clarify the definition of PRO and its scope of application in drug registration research, the general principles for PRO measurement especially the development and use of scales, quality control of PRO data collection, matters requiring attention on PRO data analysis and interpretation and points of communication with regulatory authorities.

This guideline is applicable to clinical research that use PRO as an endpoint indicator to support drug registration, including clinical trials and real-world studies.

II. DEFINITION OF PATIENT REPORTED OUTCOMES

The PRO is defined as: any assessment outcome from the patient's own

disease and corresponding treatment experience that is directly reported by the patient and is not modified or interpreted by others.

PRO emphasizes patient self-reporting. When patients do not have or lose the ability to self-assess, it may be necessary to complete the recording of PRO by their guardian or representative appointed by the guardian, but proxy bias should be fully assessed at this time.

The scale is the most used instrument for PRO measurement and mainly used for subjective measurement, such as pain, quality of life, etc., but the existing scale cannot solve all subjective measurement problems, such as certain symptoms (such as nausea) or symptom groups. There are two methods of PRO data collection in paper and electronic. The use of electronic means to record PRO is called electronic patient-reported outcome (ePRO).

III. DEVELOPMENT, TRANSLATION, AND IMPROVEMENT OF PATIENT REPORTED OUTCOME INSTRUMENTS

In clinical research, once it is determined to use a scale to measure PRO, if there is no scale suitable for the research, it needs to be developed specifically for the research purpose; if there is a recognized Chinese scale suitable for the research , it can be used directly after obtaining the copyright; if there is a recognized foreign language scale suitable for the research, it needs to form a formal Chinese version before used; if the existing scales are not completely suitable for research, they need to be modified. How to choose a scale that is more suitable for the proposed research project from the developed scales needs to consider its rationale and feasibility.

A. Development of a PRO Instrument

The development of the PRO scale should reflect the perspective of patients, with emphasis on the clinical value of the scale, including the

pertinence of efficacy evaluation, the interpretability of clinical significance and the guidance for treatment decision-making. The development process of the scale is shown in Figure 1. The development of the scale is usually used for effectiveness evaluation, but it can also be developed for important safety events. The principle and process are the same.

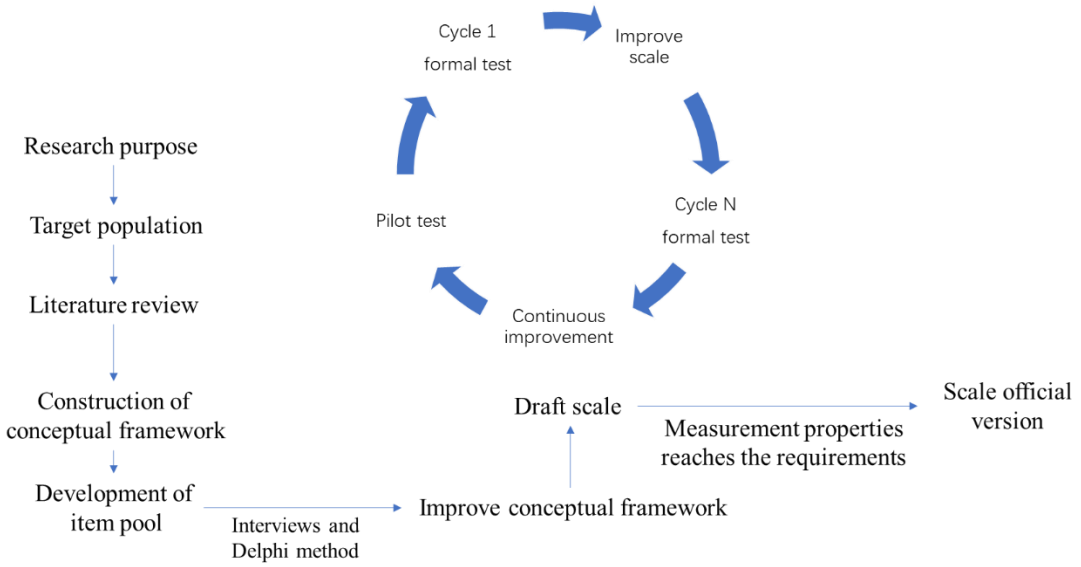

Figure 1. The schematic diagram of the scale development process

1. Construction of Conceptual Framework

The conceptual framework of the instrument composed of primary, secondary, and tertiary level. Primary or secondary level are commonly used in clinical research. The scale of the primary level structure includes single-item scales (such as the visual analog pain scale) and multi-item scales (such as the simplified xerostomia inventory, SXI). Take the secondary level scale as an example to illustrate.

The first level of the secondary structure scale is dimension, and the second level is item. The preliminary formation of the conceptual framework of the scale is generally based on literature review, specialist knowledge and experience, patient interview and necessary research. The number and

naming of the dimensions are based on the understanding of the research content. The number of items and item content under each dimension are used to reflect the connotation and importance of the dimension to which it belongs. When each item is equal in weight, the number of items under the dimension reflects the importance of the dimension.

2. Development of Item Pool

The underlying structure of the scale is the item, which reflects the specific content of the question, while the dimension is conceptual. For the subsequent item design, it is necessary to establish a pool of items as rich as possible. The source of items can be all possible channels, including literature, patient and/or specialist interviews, research and development platforms in relevant fields, research reports, developer design, etc.

Item design is one of the core contents of scale development. If the item pool is sufficiently rich and mature, most items are generally obtained from the item pool, but there will also be some items designed by the developer. In the statement of the question, closed questions should be used wherever possible, one should avoid ambiguous words, questions with dual meanings or tendentious guidance, as well as double negative statements, negative expression, and questions that the patient is unwilling to answer; at the same time, avoid ceilings or floor effect of the response, as well as asking more than two questions at the same time for one item, etc. In terms of patient understanding, it should use plain words as much as possible and the requirements for cultural level should not be too high. (For example, reading ability required for primary school graduation education)

3. Response Option Types

Response option types include two-category scale, grade scale (e.g. Likert scale), continuous scale (e.g. visual analog scale, VAS), pictorial scale and other types. Among them, the 5-level Likert scale is the most used. The

choice of response option type should be based on the best measurement performance of the scale.

4. Interviews

After the developer has initially formed the conceptual framework of the scale, it is first necessary to conduct specialist/patient interviews and/or surveys, and adjust the conceptual framework based on specialist feedback. Patient interview can help to further ensure the validity of the content of the PRO scale and reflect the needs and opinions of patients. The main purpose of the specialist survey is to verify the rationality of the structure, the accuracy of the item expression, the feasibility of the response, and the weight of the item or dimension. How to assign the weight of each dimension and item is the most critical in scale development. The specialist investigation method is usually implemented in more than one round. Especially when making decisions about weighting, specialist interviews will continue until a relatively concordance is reached.

5. Pilot Test and Formal Test

After improving the initial conceptual framework by comprehensively considering specialist opinions, the initial test version of the scale is formed, and then it needs to form the field-testing version according to the results in pilot test in a small number of target populations. A field test is carried out in the target population in which the sample size needs to be estimated according to the parameters of the pilot test, and the draft scale is adjusted according to the corresponding test results. The number of cycles of the formal test depends on whether the scale measurement properties reach the requirements.

6. Verification of the Conceptual Framework

Both pilot test and formal test are the process of verifying the conceptual framework. Fitness evaluation of conceptual framework is mainly based

on its measurement properties, including reliability and validity.

(1) Reliability: reliability refers to the ability to yield consistent measurement obtained under similar conditions. The commonly used reliability of the PRO scale includes test-retest reliability, internal consistency reliability, and inter-rater reliability. The test-retest reliability is used to evaluate the repeatability of the scale. The correlation coefficient between the initial test and the retest should not be too low. The internal consistency reliability is used to evaluate the internal consistency of the scale. Cronbach's coefficient alpha is commonly used to examine internal consistency with 0.70 as an accepted minimum. Inter-rater reliability is usually evaluated by intraclass correlation coefficient, ICC. Some literature report that ICC consistency can be divided into <0.4 means poor; 0.4-0.75 means fair-good; >0.75 means very good.

(2) Validity: validity refers to evidence that the scale measures the concept of interest as intended. An instrument should have sufficient validity and reliability. High reliability does not mean high validity (e.g., Symptoms of Major Depressive Disorder Scale, SMDDS, has high reliability and validity when used to measure major depression disorder, maybe has high reliability but low validity when used to measure mania). If a scale is not reliable, it probably is not valid.

There are many types of validity of the instrument, and 3C is more commonly used, namely content validity, criterion validity and construct validity. The content validity is mainly based on specialist knowledge and experience as well as patients' subjective opinion to judge whether the dimensions or items of the scale are reasonable, and whether they can correctly reflect the content that intended to measure. Criterion validity refers to correlation between the developing scale and the gold standard. Since the gold standard usually does not exist, and if it exists, the

significance of the developing scale is limited (only when the developing scale has great convenience, etc.), so there are few applications. Construct validity often uses exploratory factor analysis (EFA) or confirmatory factor analysis (CFA) to assess the consistency between the structure generated by the observational data and the conceptual framework.

Except 3C concepts, another important measurement property of validity is the ability to detect change, also known as responsiveness, which is the ability to sensitively reflect changes in patient outcomes (e.g., changes before and after intervention, responses to different interventions, etc.).

7. Writing an Instruction Manual of the Scale

In order to ensure the correct use of the scale, an instruction manual for the scale should be written. The instruction manual of the scale includes, but is not limited to:

- Target population
- The complete scale structure including the introductory words
- The weight of dimensions and items, the scoring rules
- Measurement properties
- The definition of validity response
- Handling missing data
- Recall period (If any)

B. Language Translation and Cultural Adaptation for Scales Used for PRO Measurement

If the original scale used for PRO measurement in clinical research is in a foreign language, it usually needs to be translated into Chinese before it can be applied. In addition, if one or several items of the original scale are incomprehensible or difficult for patients to cooperate effectively due to cultural differences, it will also involve cultural adaptation issues. Whether the translation and/or cultural adaptation of the scale is appropriate should

be measured based on whether the measurement properties of the scale after translation and/or cultural adaptation is similar to that of the original scale. The translation and/or cultural adaptation of the instrument scale can be carried out according to the following steps:

1. Preparation Stage

- Consult all relevant information about the scale development.
- Set up a multi-disciplinary translation team composed by English-Chinese translation specialists and medical professionals, etc.
- Establish a communication channel with the scale developer to obtain the license for the latest version of the scale and understand the purpose of scale development better by communication in order to translate accurately.

2. Forward Translation

Two or more translators independently translate the original language version of the scale into a Chinese version, and then synthesize each translated manuscript to form a Chinese draft.

3. Back Translation

- The Chinese draft is translated back to original language by professional translators who are native speakers of the original language and proficient in Chinese.
- Compare source and back-translated versions to identify discrepancies in the back-translations. If there is a big discrepancy, the Chinese translation needs to be further revised.
- When the discrepancy between source and back-translated versions reaches an acceptable level, the Chinese draft is formed.

4. Cultural Adaptation

If there are individual items in the scale that are not suitable for local culture, they need to be adapted. Whether the adjustment result is

satisfactory or not should be judged based on the similarity of the measurement properties of the adapted scale to the source.

5. Verification of Chinese Draft

The Chinese draft is used to conduct cognitive interviews with patients in the target population, evaluate the comprehensibility of scale items and the cognitive degree of patients, and conduct quantitative tests on the scale performance. If the measurement properties of the scale are similar to the original version, the Chinese version can be finalized. If the gap is large, the Chinese draft needs to be further improved until the measurement properties meets the requirements and the Chinese final version is formed.

6. Translation Report

After the official Chinese scale is formed, complete the translation report, which records the entire translation process, the measurement properties of translated scale, the instructions manuals. If necessary, you can declare intellectual property rights.

C. Improvement of Scales Used for PRO Measurement

When the existing scale is not completely suitable for the research, it should be used after improvement. For example, if data analysis of early clinical trial (such as phase II) shows that the scale does not meet the reliability and/or validity required of the research, it is necessary to improve the scale or develop a new scale. The scale should be tested again before phase III is launched. To ensure that the scale used in the phase III trial has sufficient reliability and validity.

IV. SELECTION AND EVALUATION OF A PRO MEASUREMENT SCALE

As a PRO measurement instrument, the scale should have good measurement properties and should be both reliable and valid. It is very important to choose an appropriate scale for PRO measurement that is

suitable for the research project to be carried out. Combining scientific and feasibility, it is recommended to focus on the following points:

A. Applicability of the Scale

Considering the construct of the scale, attention should be paid to that whether its conceptual framework satisfies the purpose of scale development and the target population, and the target population of the research should be consistent with the applicable population of the original scale.

B. Standard Documents and Systems

Standard scale-related documents or systems, including but not limited to explanatory documents (especially the interpretation of scale scores), user manuals, standard formats for data collection, important reference data (used for sample size estimation during design), etc.

C. Development Process

Whether the purpose of using the scale is clearly defined, whether the development process is strictly standardized, whether the structure of the scale (dimensions and items and their weighting) is reasonable, and whether the published results are detailed.

D. Authoritative

Whether the developed scales are published in peer-reviewed journals, whether they are widely cited and applied, and whether they are recommended by the guidelines.

E. Language and Culture

Whether the validity verification of the scale considers different educational, cultural, and ethnic backgrounds; whether the new language version has undergone a standardized translation and back translation process and verification. The measurement properties of the scale after translation and/or cultural adaption should be similar to the original scale.

F. Verification

Whether it is verified by a large enough sample size, whether the item design and weighting are reasonable, whether it has sufficient reliability and validity.

G. Feasibility

The feasibility of using the scale includes but not limited to the feasibility of the implementation process, the problem of overlapping of items when using multiple scales, etc. If the respondent burden is too heavy, it may lead to increased absence and rejection of responses, which will reduce the quality of PRO data. Factors that increase the respondent burden include: too much content in the scale, high repeatability of the content, the selection of multiple scales at the same time and one or some of the scales are of little significance, the scale interface design is not convenient for reading, the items involve privacy that is not easy to answer, and the item design is unreasonable, etc.

V. CONCERNS WHEN USING PRO INSTRUMENTS IN CLINICAL RESEARCH

A. A Framework of Estimand

The criteria and methods for constructing the estimand framework proposed in ICH E9 (R1) are also applicable to clinical research with PRO as the endpoint. The estimand framework needs to be clearly defined in the protocol and statistical analysis plan (SAP).

B. Setting the PRO as the Endpoint of the Clinical Research

If the clinical research select PRO as the primary or key secondary endpoint, the reasons and basis should be explained with factors such as research objectives, the disease mechanism, drug action mechanism, and clinical positioning of the target indication. Regarding PRO as the primary or key secondary endpoint, attention should be paid to the following issues:

- It needs to have sufficient basis and consistent with the purpose of the research;
- If the double-blind design is not adopted, it will produce a greater risk of subjective evaluation bias, which should be avoided with extreme care.
- The observation period should be long enough to reflect the clinically significant changes in PRO;
- The overall type I errors should be controlled;
- When determining sample size, full consideration should be taken that the expected difference and should have clinical significance at least.

The selected PRO should reflect the patient's perception of the effect of the drug. The effect of the drug is not limited to its efficacy, but can also be a change in safety, tolerance or impact on quality of life, etc. A reasonable selection of the PRO will help the research better reflect the patient's experience and make drug research and development follow the concept of Patient-Focused Drug Development.

C. Explanation of the Scale in the Research Protocol and Report

When the PRO measured by the scale is used as a primary or key secondary endpoint, it should be explained in the research protocol, including but not limited to: the rationality of the selection and use of the scale; if necessary, a brief introduction to the development and application of the scale, especially for some scales with less application; the evaluation methods and indicators of scale measurement properties; the collection and quality control of scale data; the analysis method of scale data; detailed instructions and training plans for the use of scales.

The clinical research report should include but not limited to: the collection of scale data (effective response, missing, etc.); the measurement properties (e.g. reliability, validity) of the scale used in the research, when it is

significantly different from the original scale, should analyze the specific reasons for the differences and evaluate the potential impact of the differences on the research conclusions; the detailed analysis results of the scale data and the corresponding reasonable explanations.

D. Valid Response

PRO measurement may be missing, negative response (such as selecting a fixed level in the response options of 5 level Likert item), which will distort the data collected from the scale. Therefore, the definition of effective response should be specified in the introduction of the scale. For example, a certain scale stipulates that if more than 15% (different scales have different definitions) of the items are unanswered, or if all item responses are same (e.g. "very good") are regarded as invalidity responses of the patient. The protocol and statistical analysis plan need to specify the judging criteria for valid responses and explain the reasons in detail. If the answer is judged as an invalidity response, it will be treated as missing data. In some cases, in addition to considering whether the entire scale is validity in response, a certain dimension of the scale may be regarded as a key variable. At this time, it may be specified in advance whether the response of the dimension is valid or not.

E. Missing Data

It is common for PRO data, especially the data measured by the scale, to appear missing. Therefore, it is necessary to strengthen quality control during the implementation of the research and reduce the missing as much as possible. For items in multi-dimensional scales, imputation method is usually adopted to deal with the problem of missing data. The specific method is given priority to the method provided by the specification of the original scale, followed by the commonly used method in the literature report, and determined by the exploratory analyses of current research data

(usually completed in the exploratory study). If missing data is not filled, except for the situation that too much missing data will make the response invalid, the entire scale and the scores of each dimension need to be adjusted according to the specification of the original scale or the rules defined in the protocol when the score of the item is missing. In clinical trial design, a reasonable statistical analysis strategy should be developed for missing data.

F. Multiplicity Issues

When PRO is a primary endpoint or key secondary endpoint, multiplicity issues will be involved. For the general handling principles, please refer to the *Guideline on Multiplicity Issues in Clinical Trials (Trial Version)*. Sponsors need to prescribe strategies and multiplicity adjustment methods for multiplicity issues in the protocol and statistical analysis plan. If certain dimensions of the PRO scale have important clinical significance and are listed as key secondary indicators in the protocol (the sponsor intends to claim the specific benefit in the instructions), it will also involve multiplicity issues, and the design needs to consider the control of overall type I errors.

Due to the multi-dimensional and multi-item characteristics of the scale, in addition to focusing on the analysis of the overall score of the scale, the analysis of each dimension and item is also necessary. Broadly speaking, it involves multiplicity issues, but if these dimensions and items are not set as the primary or key secondary endpoint, or not claim the specific benefit in the instructions, the multiplicity adjustment will not be necessary.

G. Interpretation of Results

The interpretation of the results of the PRO based on scale is the same as other endpoint indicators used to assess the benefit of treatment, and the positive results must have both clinical and statistical significance.

Minimum clinical important difference (MCID) is usually used to define the threshold of clinical significance. For example, when the 10-point visual analog pain scale is used to measure the degree of pain, how much of a decrease in scores after intervention is of clinical significance, or how much of mean difference between two groups in the changes compared to baseline is clinically significant. When determining the MCID, the relevant guidelines, specialist consensus and other recognized standards should be the first choice; if there is no recognized standard, it is necessary to communicate with regulatory agencies in a timely manner and reach a consensus, and statistical methods may provide some evidence for it.

Using statistical methods to estimate MCID, commonly used methods are distribution-based methods and anchor-based methods. The anchoring-based method is more reliable and easier to compare across different trials. This method set an external overall index (such as no improvement, slight improvement, significant improvement) according to the patient's perception of clinical significance, and then assess the amount of change in the corresponding scale score. Usually, the correlation coefficient between the overall index (rank variable) and the scale score change should be more than 0.3. Some studies believe that the correlation coefficient < 0.3 means low correlation and coefficient over 0.5 indicates high correlation. There are some other statistical methods for estimating MCID, such as methods based on mixed linear models, etc. The main approach can be determined after communication with the regulatory agencies.

H. Quality Control of PRO/ePRO

The consistency of data collection among different research centers, patients, and observers should be ensured during the implementation of the research, to improve the quality of clinical research. The protocol should at least specified but is not limited to:

- Establish standard operating procedures for quality control
- The time point and implementation sequence of PRO/ePRO data collection
- Training and guidance for relevant personnel on the use of PRO/ePRO instruments, including methods and standards for judging the completeness of the scale, and the time and method of data filling, storage, and transmission, etc. Make them fully understand the purpose of using the scale, the specific content in the manual, and the quality control in the data collection process of the scale
- PRO/ePRO data management plan

In addition, clinical research using PRO/ePRO requires more continuous and active on site monitoring to ensure the completeness and accuracy of PRO/ePRO data collection.

I. Application of PRO/ePRO in Real-World Study

In real-world study, the use of PRO/ePRO is usually limited to prospective studies, such as prospective observational studies or pragmatic clinical trials. The specific method of management or curation of collected PRO/ePRO data, please refer to *Guideline on Using Real-World Data to Generate Real-World Evidence (Trial)*.

VI. ELECTRONIC PATIENT REPORTED OUTCOMES

A. ePRO Measurement

Compared with paper-based PRO, ePRO has obvious advantages in the efficiency, real-time, flexibility, compliance, security, and patient privacy protection during data collection. The disadvantages of the ePRO is that some patients may have difficulty operating electronic devices, particularly the elderly, the young, and those with illnesses that limit their ability to operate such devices.

Interactive voice response systems (IVRS) and screen-based reporting devices are two common methods of ePRO data collection. IVRS features automatic calling, using pre-recorded questions and answer option scripts, and allows patients to use keystrokes to record responses, and the data is directly stored in the central database. The screen-based reporting system can be installed on the patient's own electronic devices, such as smartphones, tablets, computers, and even wearable medical devices, also known as Bring Your Own Device (BYOD). Patients can visit the ePRO website or software on the device, choose and record the answer according to themselves situation.

The ePRO system can be linked to an electronic medical record system (EMR) or an electronic data capture system (EDC) to form a complete data stream at the individual level. The time recording function can effectively prevent and identify behaviors that affect data reliability, such as response backfill or response in advance, etc. The remote monitoring function helps researchers and data managers to conduct online data management in real time, to mark questionable data, and to respond to subjects in time.

B. General Concerns When Using ePRO

In clinical studies for drug registration purposes, ePRO instrument, data collection and data management should follow the basic requirements of guidelines related to data management in drug clinical trial, electronic data collection, and curation of real-world data.

The data collection method of ePRO is based on network platform which is different from that of paper-based PRO. The data of ePRO is usually uploaded to an online data collection center for users' comprehensively management, data storage, monitoring and exporting. Therefore, the investigator's authority to maintain and save original electronic data should be guaranteed and research institutions have original documentation for

sponsor inspection and regulatory verification. Using ePRO measurements should according to the following:

1. Researcher's Administrative Authority

The investigator’s administrative authority to maintain and check the accuracy and authenticity of source data of ePRO should be satisfied. The investigator may capture any data changes and modifications after the ePRO data is uploaded through the measurement devices, avoiding the sponsor or the third party's sole control of the original ePRO data acquisition/management system. ePRO source data refers to the records originally recorded by the ePRO system and stored in the database. If the ePRO system original records are directly imported into the EDC system and stored in the eCRF, the original eCRF is the source data.

2. Data Security Management System and Access Control Mechanism

Encryption technology is used to ensure the integrity, confidentiality, and transparency of the data in the collection, extraction, transmission, and storage process. Any individual or organization should be prevented from modifying the original data, deleting adverse events reported by patients, high-risk warnings, and etc. A corresponding data access control mechanism should establish to avoid the risk of unplanned unblinding.

3. Data Backup

Avoid the risk of data damage or loss during the research, and the inability to reconstruct or verify the source data.

4. Data Storage

Research institutions and investigators should save electronic source data and documents, ensure regulatory investigator to inspect, verify, and copy the data at the clinical study site during an inspection.

If the analysis of the research data finds that there is a big difference between the measurement properties of the ePRO scale and the original

paper-based scale, the potential problems in the implementation of the ePRO scale should be considered and corrected. The ePRO instrument based on item response theory (IRT) can select the next item based on the answer of the previous item through computerized adaptive testing (CAT) technology. The response burden of patients can be reduced by reducing the number of items, but reducing items should meet the premise of ensuring the content validity of the scale. Sponsors using such ePRO instruments need to submit relevant materials such as the construction of conceptual framework, the designing and screening process of item bank, the rules of item selecting program, interpretation of results, etc.

VII. COMMUNICATION WITH REGULATORY AUTHORITIES

When sponsors plan to use PRO/ePRO as the primary endpoint or key secondary endpoint of a confirmatory study, they should communicate with regulatory agencies in time. Communication issues include but not limited to background of the target disease, the reasons and basis for choosing PRO as primary or key secondary endpoint, the type of research design, confirmatory conceptual framework and instruction manual of the developed scale (if any), evidence of PRO/ePRO instrument modification and/or cultural adaptation (if any), evidence of reliability and validity, specification of minimally important differences, quality control of implement process, etc. Before communication, the sponsors should provide the regulatory agency with research protocol containing PRO/ePRO statistical analysis considerations and PRO/ePRO related materials. During the trial, if major adjustments had been made to the clinical trial protocol due to changes of the PRO/ePRO, timely communication should be made with the regulatory agency.

REFERENCE

[1] Acquadro C, Berzon R, Dubois D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001.Value Health. 2003; 6(5): 522-531.

[2] Bukhari M. "PROMs vs. PREMs (Patient-Reported Experience Measures)."; Patient Reported Outcome Measures in Rheumatic Diseases. Ed. Miedany YE. London: Springer, 2016; 405-417

[3] Byrom B, Watson C, Doll H, et al. Selection of and Evidentiary Considerations for Wearable Devices and Their Measurements for Use in Regulatory Decision Making: Recommendations from the ePRO Consortium. Value Health. 2018; 21(6): 631-639.

[4] Calvert M, Blazeby J, Altman DG, et al, CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013; 27; 309(8): 814-822.

[5] Cohen J. A power primer. Psychological Bulletin 1992; 112(1): 155–159

[6] Coons SJ. ePRO systems validation: clearly defining the roles of clinical trial teams and ePRO system providers. Value Health. 2013; 16(4): 457-458.

[7] Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Research Practices Task Force report. Value Health. 2009; 12(4): 419-29.

[8] Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J Off J North Am Spine Soc. 2007; 7: 541–546.

[9] Doward LC, Gnanasakthy A, Baker MG. Patient reported outcomes: looking beyond the label claim. Health Qual Life Outcomes. 2010; 8: 89.

[10] EMA. Reflection paper on the regulatory guidance for the use of health relate quality of life (HRQL) measures in the evaluation of medicinal products. 2005.

[11] EMA. Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. 2010.

[12] EMA. Reflection paper on the use of patient reported outcome (PRO) measures in oncology studies. 2014.

[13] Fayers PM, Machin D. Quality of Life: The assessment, analysis and reporting of patient-reported outcomes (3rd Edit). John Wiley & Sons, Ltd. 2016.

[14] Ferreira ML, Herbert RD, Ferreira PH, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol. 2012; 65: 253–261.

[15] FDA. Clinical outcome assessment (COA) compendium. 2021.

[16] FDA. Clinical outcome assessment (COA) qualification program https://www.fda.gov /drugs/drug-development-tool-ddt-qualification-programs/clinical-outcome-assessment-coa-qualification-program

[17] FDA. Guidance for industry: Assessing COVID-19-Related Symptoms in Outpatient Adult and Adolescent Subjects in Clinical Trials of Drugs and Biological Products for COVID-19 Prevention or Treatment. 2020

[18] FDA. Guidance for industry: Electronic source data in clinical investigations. 2013.

[19] FDA. Guidance for industry: Patient-Reported Outcome Measures: use in medical product development to support labeling claims. 2009.

[20] FDA. Patient-Focused Drug Development: collecting comprehensive and representative input. 2020.

[21] FDA. Roadmap to patient-focused outcome measurement in clinical trials. 2015; https://www.fda.gov/media/87004/download

[22] FDA: Plan for issuance of patient-focused drug development guidance. 2017.

[23] FDA. Upper facial lines: developing botulinum toxin drug products. 2014.

[24] Fiero MH, Pe M, Weinstock C, et al. Demystifying the estimand framework: a case study using patient-reported outcomes in oncology. Lancet Oncol 2020; 21: e488–94

[25] Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics, 1975; 31: 651-659.

[26] Fox MW, Onofrio BM, Onofrio BM, et al. Clinical outcomes and radiological instability following decompressive lumbar laminectomy for degenerative spinal stenosis: a comparison of patients undergoing concomitant arthrodesis versus decompression alone. J Neurosurg. 1996; 85(5):793-802.

[27] Hong K, Majercak KR, Villalonga-Olives E, et al. Patient-reported outcomes in breast cancer FDA drug labels and review documents. J Patient Rep Outcomes. 2021; 5(1):36.

[28] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10:407–415.

[29] Lawrance R, Degtyarev E, Griffiths P, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? Journal of Patient-Reported Outcomes. 2020; 4(1):68.

[30] Ly JJ, Crescioni M, Eremenco S, et al. Training on the use of technology to collect patient reported outcome data electronically in clinical trials: best practice recommendations from the ePRO Consortium. Ther Innov Regul Sci. 2019; 53(4): 431-440.

[31] Mokkink LB, Terwee CB, Knol DL, et al. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement Instruments. BMC Med Res Methodol. 2006; 6: 2.

[32] Walters S. Quality of life outcomes in clinical trials and health-care evaluation: A practical guide to analysis and interpretation. John Wiley & Sons, Ltd. 2009.

[33] Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. Value Health. 2005; 8(2): 94-104.

[34] Guideline, ICH. "Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials." E9 (R1). Step 4 (2019): 20.

GLOSSARY

Ability to Detect Change: The measurement instrument detects the ability of the PRO measurement score to show differences with the measurement conditions (before and after the intervention, different interventions, different populations, etc.).
Adaptation: Any changes to the scale based on consideration of language and cultural differences between races. This change does not affect the structure of the PRO scale. But a small part of the content will be adjusted to fit another mode, language or population. The adaptation study is to validate the measurement properties of the PRO scale in a target environment or language.
Concept: Or Concept of Interest (COI). At the regulatory level, the concept is state or experience of an individual in clinical, biological, physiological, and functional captured or reflected by the PRO scale. At the PRO level, the concept represents the patient's functions or feelings related to their health or treatment.
Conceptual Framework of a Scale: The conceptual framework of the scale is generally based on the researcher's access to literature, specialist knowledge and experience, and necessary research. The number and naming of the dimensions are based on the understanding of the research content. The number of items and item content under each dimension are used to reflect the connotation and importance of the dimension to which it belongs. For example, when each item is equal in weight, the number of items under the dimension reflects the importance of the dimension.
Construct Validity: Evidence from whether the structural relationship between the items, dimensions, and the concept to be expressed in the PRO scale generated by the observational data is consistent with the theoretical conception of the scale development.
Content Validity: Evidence from qualitative research based on specialist knowledge to verify whether the scale can measure what it wants to measure.
Criterion Validity: The extent to which the scores of a PRO instrument are related to a known gold standard measure of the same concept. For most scales for PRO, criterion validity cannot be measured because there is no gold standard.
Cronbach's alpha: A indicator used to examine internal consistency of a scale.
Dimensions: The primary structure (secondary structure scale) or primary and secondary structure (tertiary structure scale) of the scale that represented a certain aspect (concept) of the scale. A dimension consists of

one or more items.

Instrument: A means to capture data (i.e., a scale) plus all the information and documentation that supports its use. Generally, that includes clearly defined methods and instructions for administration or responding, a standard format for data collection, and well-documented methods for scoring, analysis, and interpretation of results in the target patient population.

Item: An individual question, statement, or task (and its standardized response options) that is evaluated by the patient to address a particular concept.

Minimum Clinical Important Difference (MCID): Minimum clinical important difference (MCID) is usually used to define the threshold of clinical significance. For example, when the 10-point visual analog pain scale is used to measure the degree of pain, how much the score drops before and after intervention is of clinical significance, or how much the average score drops compared to the baseline between the two groups is of clinical significance.

Patient-Focused Drug Development (PFDD): A systematic approach to ensure that patients' experiences, opinions, needs and priorities can be captured and effectively integrated into drug development and evaluation during the entire life cycle of a drug.

Patient-Reported Outcome (PRO): Any assessment outcome of the patient's own disease and corresponding treatment experience that is directly reported by the patient and is not modified or interpreted by others.

Quality of Life (QoL): A general concept used to assess overall situation of health as reflected in all aspects of life.

Recall Period: The period of time patients is asked to consider in responding to a PRO item or question. Recall can be momentary (real time) or retrospective of varying lengths. Recall period should not be too long, such as less than a week.

Reliability: The ability of a PRO instrument to yield consistent measurement obtained under similar conditions.

Symptom: Any subjective evidence of a disease, health condition, or treatment-related effect that can be noticed and known only by the patient.

Treatment Benefit: The effect of treatment on how a patient survives, feels, or functions. Treatment benefit can be demonstrated by either an effectiveness or safety advantage. For example, the treatment effect may be measured as an improvement or delay in the development of symptoms or as a reduction or delay in treatment-related toxicity. Measures that do not directly capture the treatment effect on how a patient survives, feels, or functions are surrogate measures of treatment benefit.

Validity: The ability of the PRO scale measurement achieves the expected measurement purpose.