PROBE RESPONSIVENESS

EVALUATING THE RESPONSIVENESS OF THE PATIENT REPORTED OUTCOMES, BURDENS AND EXPERIENCES (PROBE) QUESTIONNAIRE

By VICTORIA ZUK, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

MASTER OF SCIENCE (2018) McMaster University

Faculty of Health Sciences – Health Research Methodology Hamilton, Ontario

TITLE: Evaluating the Responsiveness of the Patient Reported Outcomes, Burdens, and Experiences (PROBE) Questionnaire

AUTHOR: Victoria Zuk, B.Sc.

SUPERVISOR: Dr Alfonso Iorio, MD, PhD, FRCPC

NUMBER OF PAGES: xiii, 61

LAY ABSTRACT

This project hopes to identify the responsiveness of the Patient Reported Outcomes, Burdens, and Experiences (PROBE) Questionnaire. The responsiveness of a questionnaire is its ability to detect a change in health status when one has occurred. In order to measure whether PROBE can detect these changes, participants living with hemophilia A or B will be asked to fill out the questionnaire, as well as a few questions aimed at determining if their quality of life has changed, after they have a bleed or a surgery, as well as after 6 months. Collecting this information will help us understand how much the PROBE score needs to change in order for patients to consider a small but important change in health to have occurred. This will help with interpreting the PROBE score, which could then be used in research or in hemophilia clinics across Canada.

ABSTRACT

BACKGROUND. The study of patient reported outcomes (PROs) has seen an exponential increase in recent years. In order to be useful in practice, PRO questionnaires should be evaluated for validity, reliability, and responsiveness. Responsiveness, which assesses a questionnaire's ability to capture changes in quality of life (QOL) when they occur, has not formally been evaluated in hemophilia-specific questionnaires.

PRIMARY OBJECTIVE. To evaluate the responsiveness of the Patient Reported Outcomes, Burdens, and Experiences (PROBE) questionnaire in individuals living with hemophilia A or B following events of interest.

SECONDARY OBJECTIVES. To evaluate the responsiveness of PROBE over periods in which no events occur. To explore the use of regression analysis in aiding interpretability. To assess the presence of response shift in the study population.

METHODS. Participants will be asked to complete PROBE, as well as questions indicating changes in QOL, following a bleed or surgical intervention, and every 6 months. Responses will be evaluated using anchor-based and distribution-based approaches.

OUTCOMES. Minimally important differences (MIDs) and minimally detectable changes (MDCs) will be calculated, graphically represented, and compared to determine a single or small range of MID values.

STUDY IMPLICATIONS. Understanding responsiveness will provide increased interpretability of PROBE scores. Using an MID value, one can be confident that a change in PROBE score greater than the MID is beyond measurement error and indicates

a change in QOL. This will allow for the use of PROBE in future research trials of drug effectiveness and can offer patients' perspectives on their changes in QOL when switching to novel therapies. In addition, physicians may be able to use PROBE as a method of tracking and better understanding changes in their patients' health statuses in the clinical setting.

ACKNOWLEDGEMENTS

This thesis could not have been completed without the help of some incredible people that I have had the pleasure of knowing. For that I am grateful.

A sincere thank you to my thesis committee- Dr. Alfonso Iorio, Dr. Pasqualina Santaguida, and Dr. Lehana Thabane for the insight and suggestions that have shaped this thesis into what it is. An additional thank you for the patience throughout this project and for continuing to believe in me.

To my family and friends who have been an incredible source of strength throughout this degree, I am honoured and humbled knowing I have such an incredible support system in my life. Thank you.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF ABBREVIATIONS

DECLARATION OF ACADEMIC ACHIEVEMENT

I, Victoria Zuk, declare this thesis to be entirely my own work. The following document was developed, prepared, and written by me, with the guidance of my supervisor, Dr. Alfonso Iorio, and my thesis committee. This document has not been submitted for publication for a higher degree at another institution.

1.0 BACKGROUND

1.1 Patient Reported Outcomes

Patient reported outcomes (PROs) are measures that are reported by patients themselves.[1] A typical example is pain, as reported on a numerical scale. Clinical outcomes are instead measured by researchers, as range of motion of a joint. Both clinical and patient reported outcomes can be further defined as patient relevant (as death, or capacity to work) or weak, surrogate (as a laboratory measurement, or self-define knowledgeability). The study of PROs has seen an exponential increase in publications, and has been increasingly used in guidelines, as outcomes in various studies, and is being considered for drug approvals and quality improvement initiatives.[2–5] PRO measures aim to collect a patient's account of their experiences living with a disease and its associated treatments without interpretation by a clinician or third party.[1,4] Many important aspects of living with a condition cannot be measured without direct patient input as they are often based on what is experienced in day-to-day life.[1] Symptom presence, frequency, and severity (e.g. pain), impact of the condition on daily activities (e.g. loss of mobility), or patient perceptions and opinions (e.g. treatment satisfaction) can all be better understood by collecting PROs.[1] PROs have provided a unique approach to gaining a more holistic understanding of diseases and their associated treatments, and can be especially helpful when used in combination with clinical outcomes.[4]

1.2 Questionnaire Evaluation

Some PROs can be simply assessed with a Likert scale, like pain on a numerical or visual analog scale.[1] However, many PROs are measured through structured questionnaires. In order to be useful in practice, questionnaires should undergo evaluation of their measurement properties, such as validity and reliability. PRO questionnaires make no exceptions. Reliability is a term used to define a questionnaire's consistency and ability to provide stable results.[6,7] Test-retest reliability examines the stability of responses from the same individual at two separate time points.[1,6,7] These time points should be set close enough as to not have a change in health status between responses, but far enough to limit memory of previous responses.[1,6] A reliable questionnaire would demonstrate stability in responses across the two collection points. Internal consistency evaluates the stability between different items in a questionnaire that aim to evaluate the same concept.[7] If responses to different questions examining the same construct are highly correlated, the questionnaire is said to have high internal consistency.

Validity, on the other hand, focuses on whether a questionnaire is evaluating what it is meant to evaluate.[1,6,7] The most basic measure is face validity, which focuses on the "face value" of the associated questions, and whether they appear to be evaluating the concept in question.[6,7] Content validity aims to determine whether the questionnaire adequately evaluates the entire range of aspects within the topic at hand.[6,7] For example, a questionnaire aimed at evaluating depression should cover all aspects of depression, including the emotional, cognitive, and physical components of depression. This often requires the input and opinion of experts in the given field.[6] Construct validity also aims

to determine whether the PRO evaluates the constructs it was designed to evaluate, but concerns itself with setting and testing hypotheses.[1,7,8] This may include evaluating internal relationships, evaluating validity across a variety of groups (e.g. different languages or cultures), or comparing the PRO in question to other related instruments.[7] Construct validity can be tested by assessing whether the PRO at hand correlates with another instrument that looks at the same construct or with real-world performance.[8] Indeed, many of these hypotheses will address other types of validity at the same time.[8] Criterion validity aims to compare the novel tool with another validated measure, often the "gold standard", in order to examine its adequacy in evaluating the concept in question.[1,6,7] Concurrent validity is similar, but instead assesses the correlation between the questionnaire being evaluated and other similar measures.[8] This can include any measure attempting to assess the same constructs and does not necessarily need to be the gold standard. Discrimination is another aspect of PRO evaluation that should be considered. Discrimination examines the questionnaire's ability to distinguish responses provided by different groups of people.[6,9] Importantly, the questionnaire should be able to discriminate between individuals with different health statuses, such as varying disease severities.[9] Those PROs that cannot discriminate between important groups may be limited in their value in clinical settings.

1.3 Responsiveness

In addition to assessing a questionnaire's reliability and validity, there are other important measurement properties that should be considered. One such characteristic is responsiveness, which aims to evaluate a questionnaire's ability to detect a change in a measure when one has occurred.[7,10] As mentioned earlier, discrimination aims to evaluate differences in scores between two individuals with different health statuses.[9] Responsiveness on the other hand, looks at the change in health status of one individual over time.[10] This characteristic has been identified using other names, including sensitivity to change or longitudinal validity, but the term responsiveness was suggested in order to avoid confusion with diagnostic sensitivity and other measurement properties.[11–13] The literature on this topic provides a variety of definitions, methods, and statistical approaches to evaluation of responsiveness.[10,11]

Both cross-sectional and longitudinal methods can be used to evaluate responsiveness.[10,14] Longitudinal methods track participants and their health statuses over time. A cross-sectional approach requires individuals to discuss their present health status with another patient and rate themselves based on the conversation.[14] This discussion provides a comparison point for participants to judge their own health status and offers investigators a method of understanding differences in quality of life (QOL). Although this technique provides a simple way to evaluate responsiveness at a single point in time, only between-person differences can be evaluated.[10,14] Given that responsiveness looks to evaluate change in health status over time, most investigators chose to evaluate within-person differences, making prospective designs more common.

When selecting a prospective design, there are two main approaches that can be used: the anchor-based and the distribution-based approaches.[14–16] Typically, these approaches are applied across a period in which participants are receiving a therapy of know efficacy as a method of changing QOL through the course of the study.[12,13] The anchor-based approach entails using an additional measure to assess how the participant is progressing. This measure is termed the "anchor".[17] This can be a PRO, such as a global health rating, or a non-PRO measure (herein termed an "objective measure"), such as a laboratory value or a clinical outcome.[14,18] Ideally, the anchor that is selected should be an accepted measure of QOL or should attempt to capture the same construct related to the change as the PRO being evaluated.[19] Nevertheless, the anchor and the questionnaire at hand should be correlated.[19] Anchors are primarily used to sort individuals into groups based on amount of change exhibited in the hopes of identifying those who have experienced a small, meaningful change, and those who have not changed at all.[14]

The most commonly applied anchor is the global health rating question, in which participants are asked to rate the amount of improvement or decline they have experienced on a scale from +7 to -7(Figure 1).[18] This anchor is not without its concerns. Unlike other anchors which can be applied at baseline and follow-up, this anchor asks for an assessment of change in the follow-up period. For this reason, global health rating questions may be subject to recall bias; individuals are unlikely to recall their baseline status to the same degree as they are able to assess their current status.[14,17,20,21] In addition, the validity and reliability of these questions remain unknown. Nevertheless,

they have previously been shown to be sensitive to changes in both directions (improvement or decline).[14]

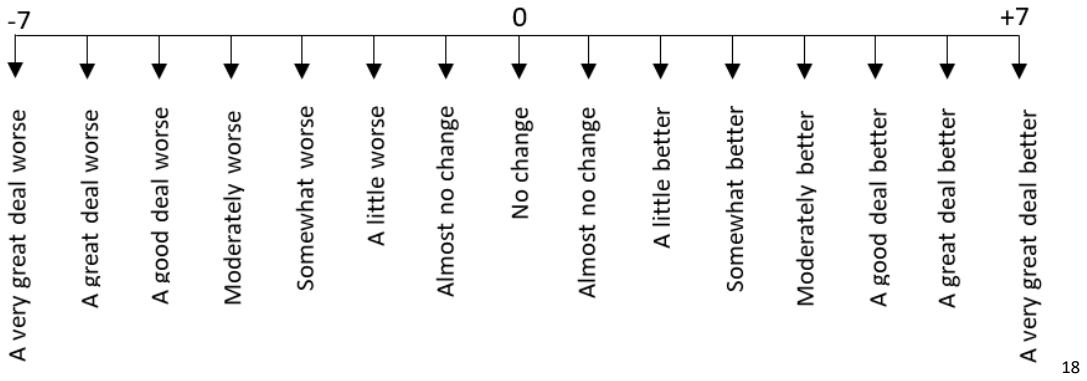

Figure 1. Global Health Rating Question

Responses to patient-reported anchors are also susceptible to response shift. Response shift occurs when an individual's values, perceptions, or standards for assessing their own QOL change over time.[21–24] With a response shift, one may observe a change in objective measures of health status despite a stable patient-reported measure, or vice versa.[22,23] This issue may lead to biased responses to the anchor, which in turn, can distort the resulting findings. Although response shift has primarily been evaluated in people with chronic diseases or those who have experienced a significant event (such as a new diagnosis), it is a concern that applies to all longitudinal PRO collection.[22] It is possible to formally assess the presence of a response shift using a variety of statistical methods. Sajobi et al provide an overview of possible methods to apply, one of which includes Oort's SEM.[22] This approach allows for the evaluation of a variety of types of response shift.

Using objective anchors that are external to the patient's perspective are sometimes preferred to patient-reported ones.[25,26] Using PROs to evaluate changes in other patient-reported measures has been criticized as only assessing concurrent validity of the tools, rather than assessing responsiveness.[25,26] That being said, objective measures may not always be possible or feasible to use. For example, there is no one blood test or score that is representative of QOL in hemophilia. The Hemophilia Joint Health Score could be used as an anchor, but unfortunately this is not representative of the entire scope of QOL. The number of days of work or school missed could be used, but the same issue arises. In order to provide a more accurate representation of QOL, physicians could be asked to assess the entire scope of the individual's case and make a judgment on whether or not a change has occurred. Unfortunately, this would require a judgment call to be made and there may not be agreement between physicians. This brings up the discussion of how much of a change represents a minimal change, as well as whether that change would be considered important.[20,25] The possibility of discrepancy in what is considered to be a "minimally important" change is concern that is often raised when applying objective anchors.[20,25] In addition, it can be argued that changes determined using objective anchors may be capturing clinically important changes and may not reflect what patients believe to be important changes in QOL.[25] For example, a clinical score may indicate a participant's life expectancy has improved after receiving a new treatment, indicating an improvement in QOL. At the same time, the individual may now be experiencing novel adverse side effects or new limitations that may result in them perceiving that their QOL has remained unchanged or has worsened.

As a method of evaluating anchor-based approaches to responsiveness assessment, minimally important differences (MIDs) are typically calculated. MIDs represent the smallest change in score that has been identified to indicate a meaningful change.[14,16] Multiple approaches to evaluation exist, including calculating correlation, using receiver-operator characteristic curves, using regression, or calculating mean change scores.[19,27] Mean change score are most commonly used in the field and can be calculated as absolute or relative scores.[28] The appropriateness of each may rely on the overall score of the questionnaire being tested.[28] For example, a 10-point change in score from 100 to 90 and from 20 to 10 may or may not represent the same amount of change in participants. In absolute terms, these both represent a 10-point change, whereas they represent a 10% and 50% change in relative score, respectively. It has been suggested to select the measure that is more closely correlated to a change in the anchor.[28]

Of note, certain approaches to assessing MIDs, including the global health rating, have been said to only addresses a "minimal" change in health status.[10] Beaton et al. have suggested that in order to truly calculate an MID, "importance" of the event must also be considered.[10] This can be done by adding an additional question asking participants to assess the importance of the change in QOL experienced.

The other approach to evaluating responsiveness prospectively is termed the distribution-based approach. Multiple distribution-based methods have been reported in the literature; the most common method being the standardized error of the measure (SEM), effect size (ES), and standardized response mean (SRM).[14–17,27] The equations for calculating these statistics have been provided in Appendix A. Given that statistics are

primarily used, these methods have been the foundation for defining a minimal detectable change (MDC) and not an MID.[14,29] Rather than representing the smallest change that is known to be meaningful, an MDC represents the smallest change that can be determined beyond measurement error.[14,30] The distinction here is that MDCs do not take "importance" into consideration, but rather focus on statistical properties.[15,17] Despite evaluating different concepts, the MDC is said to approximate the MID, although this remains to be a point of contention.[14,29,30] For this reason, researchers should recognize that distribution-based approaches may not be ideal for calculating MID, and should be used as supporting evidence when anchor based methods are available.[17,29,31]

Following this logic, it is often suggested that multiple approaches should be used concurrently to estimate a variety of possible MIDs, which can then be examined in hopes of determining a single MID for the questionnaire.[16,20] To do this, it is first recommended to graphically represent the results obtained using all approaches to facilitate their comparison.[20] Then, given that anchor based approaches have some consideration of importance of the change, results obtained from these methods should be provided with a higher weighting during triangulation, as should anchors that are more closely related to the measure of interest.[20] Again, it has been recommended that distribution-based approaches should be used as supporting information when anchor-based approaches are available.[20,29] For example, if two MID values are calculated using two anchors (one global health rating and one laboratory test) and one MDC is calculated using an SEM, the MDC would be used as a guide. Given that the MDC provides the smallest change that is detectable, if one of the MID values is smaller than the MDC and one is larger, the

larger value is likely to be correct. If both MIDs are above the MDC, the global health rating may be provided with a stronger weighting when comparing the two values as it may be more representative of QOL than a laboratory test result.

If there are many results or if the results prove to be difficult to compare, the Delphi approach can also be applied to further narrow the range of values or to agree upon a singular MID value.[20,31] This would involve gathering the opinions and insights of experts in a structured and anonymous format through multiple Delphi rounds in order to best estimate the MID value. The Delphi approach recruits experts and distributes anonymous questionnaires on the topic at hand to each of them.[32] Once responses are obtained, the opinions of fellow colleagues are distributed back to everyone on the panel.[32] The questionnaires are then filled out again, and in this iterative process, an eventually consensus is obtained.[32]

It is important to note that one true MID does not necessarily exist. The MID that is determined is context-specific, represents the population at hand, and may differ for other populations with different characteristics, such as severity of disease.[17,20,31] For this reason, one should understand the characteristics of the population that is used for establishing this measurement properties.[33]

Despite responsiveness being a fundamental property, it is sometimes overlooked when evaluating PROs.[34] Understanding whether or not a questionnaire can capture changes in a measure can help clinicians and patients better understand their health status. A responsive questionnaire could be used to help evaluate the effectiveness of treatments,

or to guide clinic visits based on any observed changes in a PRO score. With a responsive questionnaire, one could be sure that observed changes are meaningful, and not a reflection of measurement error.

1.4 Hemophilia

Hemophilia is an inherited bleeding disorder in which individuals have a decreased ability to produce clots.[35] This is the result of a lowered level of clotting proteins in the blood. Hemophilia can be classified into A or B, representing a factor VIII and IX deficiency, respectively.[34] The severity of the disease is related to the amount of residual levels of clotting factor in the bloodstream, with lowest levels relating to more severe disease.[34,35] Severity clinically translates to bleed frequency and severity, with more severe patients experiencing more prolonged or spontaneous bleeds.[36] These events tend to occur in the joints or muscles, and may cumulatively lead to deterioration, chronic pain, and/or loss of function.[37,38] This may interfere with their activities of daily living and can significantly reduce the individual's QOL.[34,37] Although hemophilia A and B share many similarities, their clinical presentation may be different, with the conditions exhibiting different frequencies and severities of events.[39–41]

For some time, the disorder could not be cured. Now, with gene therapies, a handful of individuals have demonstrated the promise of these approaches in curing hemophilia.[42,43] Until these therapies are widely available and adopted, the goal of treatment is to lessen the burden of bleeding events.[44] Individuals may take factor replacement therapies either on a regular prophylactic schedule to prevent bleeds from

occurring or on an episodic basis in response to a bleed to try to limit the duration of the event.[45,46] Research has shown that joint health can be maintained under prophylactic treatment schedules.[46]

As a method of evaluating treatment effectiveness in hemophilia, research has historically focused on non-PRO clinical outcomes such as presence of target joints (defined by ISTH as joints with 3 or more bleeds in 6 months) or number and duration of hospital visits.[35] Given that a large component hemophilia morbidity is related to the disease impact on QOL via pain, loss of function, and disability, there has been a rise in interest in the use of PROs to capture these health problems.[34,38] In fact, there are multiple questionnaires, both generic and disease-specific, that can and have been used in individuals with this disease. In terms of generic questionnaires, the EQ-5D and SF-36 are most commonly applied to assess global health status.[38] As for disease-specific questionnaires, a variety have been developed in the literature and each of these measures are PRO in format, including but not limited to CHO-KLAT, Haemo-Qol, and Hemofilia-Qol.[34] These disease specific measures have had important measurement properties evaluated (i.e. validity and reliability) to varying degrees and have their unique strengths and weaknesses. The degree of evaluation of these PROs is detailed in Appendix B. Unfortunately, many of these questionnaires were developed prior to FDA PRO guidance publication, and therefore are unlikely to have followed the rigorous guidance provided.[47] For example, despite having some patient involvement throughout the project, patients have not been consistently involved at the beginning of development or have not had large involvement in the decision-making process. Instead, experts have made a large

majority of design decision independently.[47] Furthermore, despite most questionnaires addressing internal consistency, test-retest reliability, content validity, and hypothesis testing, none have formally evaluated responsiveness in this population.[34] Many of these PROs are widely used and most of them have been translated into multiple languages, and yet, it is uncertain if they are able to detect changes in QOL over time.[34]

1.5 Patient Reported Outcomes, Burdens and Experiences (PROBE)

The Patient Reported Outcomes, Burdens and Experiences (PROBE) questionnaire was designed by patients for patients, with a complete involvement of the patient component the inception.[48] The questionnaire contains 3 sections, including general demographics, disease-specific questions, and general health problems, with the EQ-5D being collected as an additional component.[48] Of all the information collected, only the general health problems section informs score calculation for this questionnaire. Feasibility testing has demonstrated that PROBE is of low burden, with majority completing the questionnaire within 15 minutes.[48] It has since been evaluated for content validity, test-retest reliability, and other psychometric properties. PROBE was found to have excellent overall agreement, as well as good internal consistency, strong correlation with applicable EQ-5D domains and utility index score, and excellent discriminative validity.

Currently, PROBE is being collected in an anonymous manner across multiple countries and in multiple languages. As it stands, PROBE only requires the participant's country and language for completion. Participants are offered a code to continue

completion if it is interrupted, but this code is unique to the questionnaire and not the participant. If the same participant returns to complete a second questionnaire, a new code will be generated and the second response will not be linked to the first. Unfortunately, the anonymous nature of the questionnaire has not allowed for the collection of any longitudinal data on an ongoing basis. Test-retest properties were evaluated in a separate study by participants manually identifying their subsequent questionnaires. Other than these select few participants and responses, PROBE data has generally remained unlinked. This has limited the possibility of evaluating the responsiveness of the questionnaire to date.

In order to address this aspect further, it is proposed that PROBE be linked to longitudinal data, such as that collected by the Canadian Bleeding Disorders Registry (CBDR). CBDR is a database that allows prospective data collection by patients and physicians, and includes information relating to bleeds and treatment infusions. To date, CBDR exists in 26 centers across Canada, and contains records of 2032 individuals living with hemophilia A and 453 living with hemophilia B. A large subset of those enrolled in CBDR also use MyCBDR, a mobile application that facilitates data entry and collection for participants. This allows for continual tracking of treatment infusion and events, and aids in monitoring the condition. The CBDR/MyCBDR program, database, and website are hosted and administered a McMaster.

By linking these two McMaster-based initiatives, PROBE responses could be followed over time, and could be linked to bleed and treatment data. Given that we are

interested in evaluating the responsiveness of the PROBE questionnaire, this linkage would provide the perfect platform for this type of evaluation.

2.0 STUDY FLOW

The proposed study will include multiple phases, each of which will be explained in detail in this document. The initial phases have been designed to gather vital information that will inform certain aspects of the main study.

- Phase 1. A Delphi panel will be conducted to confirm the timepoints for item collection. This stage will help identify optimal times to evaluate short and longer-term effects of events on patient QOL.
- Phase 2. Once these timelines are confirmed, a subset of participants will be recruited for a pilot study. This will help identify the number of "minimally important" events that can be expected as a method of estimating an appropriate sample size. In addition, this phase will aid in ensuring the selected timepoints and study methods are feasible and appropriate for the main study.
- Phase 3. Once the timeline and sample size are confirmed, the main study will be implemented. Participants will be recruited and asked to complete the items at the given timepoints, as well as at a 6-month follow-up. Each participant will be followed for a year following baseline.

- Phase 4. Upon completion of the study, data analysis will occur. A second Delphi panel may be required depending on the complexity of the results in order to identify a single or small range of MID values.

3.0 STUDY PURPOSE

3.1 Phase 1: Delphi Study

This phase aims to determine the optimal timepoints for data collection for the main study. The timepoints for the main study require thoughtful consideration in order to capture anticipated trends in change in QOL across events. A Delphi study would provide a great platform to capture the opinions of experts in the field who may have insight on what happens in practice.

3.2 Phase 2: Pilot Study

The pilot study will serve two main purposes. Primarily, the pilot will help investigators understand the proportion of bleeds and surgical interventions that result in minimally important changes in QOL, as indicated by participants. Using this information, study staff will be able to estimate the number of individuals required to observe at least 100 meaningful bleeds and 20 meaningful surgeries throughout the study period.[49]

In addition, this phase will provide an opportunity to confirm the planned methods for the main study. For example, the pilot will allow study staff to implement the timepoints determined by the Delphi panel and determine whether the anticipated trends can be seen. The mobile application that will be used (as discussed in section 4.3.7) can also be tested to ensure ease of registration and use. Finally, the appropriateness of the selected anchors will also be tested. This will provide the opportunity to implement a more suitable anchor if required before moving forward with the main study. The planned pilot study will not only help with estimating a sample size, it will also provide an opportunity for troubleshooting before the main study is implemented.

3.3 Phase 3: Main Study

The purpose of this phase is to evaluate the responsiveness of the PROBE questionnaire. As a primary objective, responsiveness will be evaluated in Canadian individuals living with hemophilia A or B across events of interest, including bleeds and surgical interventions. As a secondary objective, responsiveness will be evaluated in the same population over a period in which no events are experienced. Regression analysis will also be assessed as a secondary objective, as will the identification of possible response shift in the study population. These secondary objectives will aid in understanding the context of the results and interpretation. In addressing these objectives, a mixed approach will be taken, in which anchor-based and distribution-based methods will be used to calculate a variety of MIDs and MDCs. An overview of the study objectives can be seen in Table 1.

Table 1 General overview of study objectives

Objective	Outcome Variable	Explanatory Variable	Hypothesis	Method of Analysis
Primary: To evaluate the responsiveness of the PROBE questionnaire following events of interest	Change in PROBE score	Change in response to global health rating	The PROBE score will increase when QOL is reported to improve and decrease when QOL is reported to deteriorate	Primary: MID calculated using mean change score
	Change in PROBE score	Change in EQ-5D score	The PROBE score will be directly correlated with the EQ-5D score	Supportive: MID calculated using mean change score
	Change in PROBE score	Smallest change beyond measurement error	PROBE will show consistency in the distribution-based methods for MDC	Supportive: MDC using ES, SEM, and SRM.
Secondary: To evaluate the responsiveness of the PROBE questionnaire in the absence of events	Change in PROBE score	Change in response to global health rating	The PROBE score will increase when QOL is reported to improve and decrease when QOL is reported to deteriorate	Primary: MID calculated using mean change score
	Change in PROBE score	Change in EQ-5D score	The PROBE score will be directly correlated with the EQ-5D score	Supportive: MID calculated using mean change score
	Change in PROBE score	-	-	Supportive: MDC using ES, SEM, and SRM.
Secondary: to explore regression analysis for responsiveness	Global health rating	PROBE score	The reported QOL will increase as the PROBE score increases	Simple linear regression
	EQ-5D			
Secondary: to assess the presence of response shift				Oort's SEM approach[50]

3.4 Phase 4: Delphi Panel

A decision regarding whether or not the final phase will be required will depend on the results obtained from the main study. If the analyzed data proves to be very spread or difficult to interpret, the Delphi panel may be required. The purpose of this second panel is to narrow down the study results to a single value or a small range of values. Accurately narrowing down this range of values will ensure proper future interpretation of changes in PROBE score.

4.0 METHODS

4.1 Phase 1: Delphi Panel

This phase will begin with the recruitment of experts in the field. Given that this phase aims to determine ideal timepoints for data collection, individuals would be considered experts if they have an understanding of the management of hemophilia A or B, as well as the recovery process following disease-specific events (i.e. bleeds and surgical interventions). Ideally, this panel will include individuals living with hemophilia A, individuals living with hemophilia B, and physicians or researchers that have been working in the field for at least 5 years. Canadian experts would be considered first, given that their knowledge and experience would reflect the anticipated study population. This phase will have a target of 20 experts and will selectively recruit to ensure a distribution of patients, physicians, and researchers.

Once recruited, the assembled panel will be contacted by a facilitator.[32] Panel members will be provided with a set of statements regarding timepoints in which individuals would experience:

- Negative impacts on QOL after a bleed (such as pain or loss of function)
- Improvement in QOL after recovery from a bleed
- Negative impacts on QOL after a surgical intervention (such as pain or loss of function)
- Improvement in QOL after recovery from a surgical intervention

Experts will be asked to rate the degree to which they agree with the statements provided by the facilitator, and once complete, the ratings will be submitted. Using the completed ratings, the facilitator will provide feedback to each individual in an anonymous manner. Experts will then have the opportunity to compare their responses with those of the group and provide updated responses in the second round of questioning. These structured steps will continue until the responses of the group converge.[32]

By implementing this phase, a timeline for data collection can be identified to suit the trends in QOL the main study aims to capture.

4.2 Phase 2: Pilot Study

Once the Delphi panel is completed, a pilot study will be implemented. This phase will follow the methodology as highlighted in the main study (section 4.3). Individuals

will be considered eligible if they meet the inclusion criteria of the main study. The pilot study will utilize the timepoints for data collection as indicated by the Delphi panel in order to help confirm their appropriateness. Unique to this phase is a short set of questions in which participants will be asked to provide insight on the study design, as well as the user experience using the mobile app for the study purposes.

4.3 Phase 3: Main Study

The main study will be implemented upon the completion of the Delphi panel and the pilot study. Vital information from these first two phases will help inform the main study.

4.3.1Study Design

A multicenter, prospective repeated measures study will be implemented. This phase of the study will take place across multiple CBDR centers across Canada.

4.3.2 Inclusion Criteria

- ≥16 years of age
- Living with hemophilia A or B
- Enrolled in or willing to enrol in MyCBDR

4.3.3 Exclusion Criteria

- Unable to read or understand English or French

4.3.4 Recruitment

In order to participate in this study, participants will be required to login to PROBE app (described in section 4.3.7) using their MyCBDR accounts. Once logged in, they will be able to access all study information. Participant consent will be received electronically prior to completing the first questionnaire.

Participants of the pilot study will be carried over to the main study provided that none of the time points or anchors are changed. If the main study remains unchanged from the pilot, these individuals will continue their enrolment and will contribute to the collection of study data.

4.3.5 Data Collection Timeline

Following consent, participants will be asked to complete their first PROBE questionnaire, which will act as a baseline. Although baseline demographics are collected in CBDR, participants will be asked to complete this section of PROBE during their baseline interaction in order to ensure this information remains accurate.

Participants will be asked to complete the PROBE questionnaire at multiple timepoints following events of interest and at regular intervals. These timepoints will be determined in Phase 1 (Delphi panel) and confirmed in Phase 2 (Pilot study). As indicated above, data will be collected at points in which participants experience:

- Negative impacts on QOL after a bleed (such as pain or loss of function)
- Improvement in QOL after recovery from a bleed
- Negative impacts on QOL after a surgical intervention (such as pain or loss of function)

- Improvement in QOL after recovery from a surgical intervention

It is possible that the selected timeline will roughly resemble the following example:

- Within 1-2 days following a bleed (acute)
- Within 7 days following a bleed (long term)
- Within 7 days following a surgical intervention (acute)
- Within 3 months following a surgical intervention (long term)

A 6-month follow-up will be added as an additional timepoint to the list mentioned. This timepoint will be required 6 months after the last completed entry; those who do not experience any events or those who filled their last questionnaire 6 months prior will be asked to complete this follow-up questionnaire.

All participants will be asked to continue the study for 1 year following their baseline encounter. Those individuals that participated in the pilot study (if applicable) will be asked to complete 1 year from their initial enrolment in the pilot study. Those individuals who are recruited during the main study will be asked to complete 1 year from their initial recruitment date.

4.3.6 Items for Collection

At each of the timepoints indicated above, participants will be asked to complete the following items:

- The PROBE questionnaire (including EQ-5D)
- A global health rating question (figure 1)
- An importance question related to an overall change in their QoL

As previously mentioned, the PROBE questionnaire includes four sections: baseline demographics, disease-specific questions, general health problems, and EQ-5D. To facilitate data collection, questions regarding baseline demographics will be asked at the first interaction but will be removed for subsequent completions.

The PROBE score will be calculated using only information provided in the general health problems section. As mentioned, the EQ-5D is also collected as part of PROBE but this information does not enter in the score calculation. This portion will act as an anchor and will be used as supportive information.

Once the questionnaire is completed, participants will be asked if they experienced a change in QOL since their last data entry point. At this point, they will be prompted to complete a 14 item global health rating score that will demonstrate the degree of improvement or deterioration experienced (from +7 to -7) (seen in figure 1).[18,27] Those who have indicated a change in QOL will be asked to indicate whether or not that change was important to them ("Would you consider this to be an important change? – Y/N"). These questions represent the primary anchor of interest.

Table 2 Timeline of items to be collected

Items for Collection	Baseline	Bleed: acute	Bleed: long term	Surgery: acute	Surgery: long term	6mo timepoint
Baseline demographics	X					
PROBE questionnaire	X	X	X	X	X	X
Global health rating		X	X	X	X	X
Importance question		X	X	X	X	X

4.3.7 The PROBE App

PROBE developers have agreed to establish a mobile app that will be ready for use prior to the implementation of this study. This app will be the primary method of data collection for this study and will allow for the collection of all necessary items.

Not only will the system be able to collect longitudinal data, it will also include functionalities aimed at reducing the burden of completion on individuals. The demographics portion of the PROBE questionnaire will be omitted in subsequent timepoints in order to limit the number of required fields. The app will also allow for offline questionnaire completion. Submitted data will automatically upload to the database once a stable internet connection is available. In order to potentially limit the amount of missing or incomplete submissions, the app will also contain a notification system which will remind participants when to complete the questionnaire. The notifications will be emailed to the participants and will display directly on the individuals' phones or tablets. Given participants will be registering bleed data on MyCBDR, the PROBE app will be able to provide notifications that are event specific as well.

4.4 Phase 4: Delphi Panel

This final Delphi panel may or may not be required, depending on the data collected and analysed from the main study. In order to aid in the process of narrowing to a singular MID value, experts will be recruited for a Delphi panel. Individuals would be considered experts if they have knowledge in the PROBE questionnaire, and particularly

in its scoring. Individuals who were involved in the development process would be considered great candidates, and could therefore include physicians, researchers, or patients. This phase will have a target of 20 experts and will selectively recruit to ensure a diverse panel.

Once recruited, the assembled panel will be contacted by a facilitator.[32] Panel members will be provided with an explanation of the study methodology, along with context for each presented value. A set of statements about the probable MID value will be sent out by the facilitator and experts will be asked to rate the degree to which they agree with the statements. Once completed, responses should be submitted. Using the completed ratings, the facilitator will provide feedback to each individual in an anonymous manner. Experts will then have the opportunity to compare their responses with those of the group and provide updated responses in the second round of questioning. These structured steps will continue until the responses of the group converge.[32]

By implementing this phase, a singular or small range of MID values can be identified.

5.0 DATA ANALYSIS

5.1 Phase 1: Delphi Panel

For this phase, data will be analyzed after each consecutive step. Study investigators will evaluate the amount of agreement between the experts. Steps will continue until consensus is reached, which for the purposes of this phase will be defined as 80% agreement.[32] Once agreement is reached, this phase will be completed.

5.2 Phase 2: Pilot Study

The pilot study aims to address a variety of questions. In order to determine an adequate sample size, investigators will determine the proportion of bleeds and surgical interventions that result in meaningful changes in QOL, as indicated by participants. Using this information, study staff will be able to estimate the number of individuals required to observe at least 100 meaningful bleeds and 20 meaningful surgeries throughout the study period.[49]

In order to confirm whether the collection points are appropriately spaced, collected data will be assessed to identify trends in scores. The PROBE questionnaire, EQ-5D, and responses to the global health rating will be assessed. It is expected that QOL will decrease in the acute timepoints and will increase across the longer timepoints. Study investigators will assess whether these trends can be seen in the collected data. This will provide an opportunity to validate the information provided by the Delphi panel and will ensure the timelines provided by experts reflect what occurs in the study scenario.

Responses to the questionnaire provided at the end of the pilot study will be assessed qualitatively. Participant opinions and responses will be evaluated to determine any areas that may require troubleshooting before the main study is implemented. Of primary concern will be questions regarding the mobile application, in order to ensure any potential issues can be resolved before moving to the next phase.

The appropriateness of the selected anchors will also be assessed. The collected PROBE scores will be compared with both the global health rating scale and the EQ-5D scores. If the calculated correlation meets a minimum of 0.3, the anchor will be deemed appropriate and will be used in the main study.[16,20] If the anchor does not reach this minimum, study investigators will consider using a novel anchor for the next phase.

Provided the timeline can capture the intended trends in QOL and there are no major issues, the study will move onto phase 3.

5.3 Phase 3: Main Study

Before primary or secondary analyses are performed, the suitability of the selected anchors will be assessed once more using main study data. This will be done by evaluating the correlation between the anchors (global health rating and EQ-5D) and the PROBE scores. The anchors will be deemed suitable if the correlation is 0.3 or greater.[16,20]

5.3.1 Primary Analysis

As a primary objective, the responsiveness of the PROBE questionnaire following events of interest will be evaluated. To determine change in health status over time, repeated measures will be collected. Differences in scores between those who experienced small and important changes and those who remained unchanged will be calculated. MIDs for improvement and deterioration will be considered separately and will be calculated using mean change scores.[15,31] The decision to use absolute or relative changes will be made based on the collected data.[28] Suitability will be determined by evaluating the correlation between the absolute or relative changes with the global health rating, as suggested by Zhang et al.[28] If either approach is suitable, absolute mean change scores will be calculated given the increased interpretability of the results.[28]

Of primary interest are MIDs calculated using the global health rating question and the importance question. Individuals who indicate a small (+/- 1, 2, or 3 on the global health rating) and important (Y on the importance question) change will be selected and compared to those who indicated no change (0 on the global health rating) in QOL. MIDs using EQ-5Dscores will be calculated in the same way, although, these MIDs will be used as supporting information. Additional supportive information will be obtained by calculating MDCs. ES, SEM, and SRM will be calculated. ES will be calculated by dividing the change in PROBE score by the standard deviation at baseline, whereas the SRM will be divided by the standard deviation of the change. These equations for calculation can be found in Appendix A.0.2, 0.5, and 0.8 will represent small, moderate,

and large responsiveness for ES and SRM approaches, whereas 1 SEM will be estimated to approximate the MID.[14,17,19]

This analysis will first consider bleeds and surgical interventions together to determine an overall MID value for the PROBE questionnaire. Subsequently, bleed data and surgical data will be separated and analyzed using the same methodology. This provides the opportunity to determine whether PROBE's ability to detect change in QOL is affected by the event of interest. For an overview of data analysis for the primary objective, please refer to Table 3.

Table 3 Data analysis plan for primary objective

Outcome Variable	Event of Interest	Explanatory Variable	Direction of Change in QOL	Comparator Participants	Method of analysis	Primary or Supportive ?
Change in PROBE score	Any event	Global health rating + importance	Improvement	No change	Mean change score	Primary
			Deterioration			
		EQ-5D	Improvement	No change	Mean change score	Supportive
			Deterioration			
	Bleed	Global health rating + importance	Improvement	No change	Mean change score	Primary
			Deterioration			
		EQ-5D	Improvement	No change	Mean change score	Supportive
			Deterioration			
	Surgical Intervention	Global health rating + importance	Improvement	No change	Mean change score	Primary
			Deterioration			
		EQ-5D	Improvement	No change	Mean change score	Supportive
			Deterioration			
	Any event	-	-	-	ES, SEM, and SRM	Supportive
	Bleed	-	-	-		
	Surgical Intervention	-	-	-		

Once these calculations are performed, a variety of possible MID/MDC values will be determined. In order to home in on a single value or a small range of values, the calculated items will be represented graphically and analyzed. As previously stated, the MIDs determined using the global health assessment will receive the highest weighting.[20] Those calculated using EQ-5D scores will receive the next highest weight, and distribution-based approaches will receive the smallest weighting.[20] If the values prove to be very different from each other and selecting a single value proves to be difficult, the Delphi approach will be used once again.[20]

5.3.2 Secondary Analysis

As a secondary objective, PROBE responsiveness will be evaluated using participants that have not experienced an event of interest. Using responses provided at the 6-month timepoint, participants who have experienced small and important changes in QOL will be identified and compared to those who have not experienced any changes. Again, MIDs will be calculated separately for improvement and deterioration.[15,31] Results from the global health assessment will be of primary interest, and those from the EQ-5D and from distribution-based methods will act as supportive information. This information will be analyzed using the same process described above, and the Delphi approach may be used for this objective as well, if needed. Table 4 provides an overview of data analysis for the secondary objective.

Table 4 Data analysis plan for secondary objective

Outcome Variable	Event of Interest	Explanatory Variable	Direction of Change in QOL	Comparator Participants	Method of analysis	Primary or Supportive?
Change in PROBE score	6-month follow-up	Global health rating + importance	Improvement	No change	Mean change score	Primary
			Deterioration			
		EQ-5D	Improvement			Supportive
			Deterioration			
		-	-	-	ES, SEM, and SRM	Supportive

Regression has also been suggested as an interesting way to evaluate responsiveness.[19] Although this has not been as widely calculated as mean change scores, this approach may warrant consideration. For this reason, regression analysis will be performed as a secondary objective. The PROBE score will be set as the independent variable, whereas the global health assessment score or EQ-5D will be set as the dependent variable. A simple linear regression model will be conducted.

In addition, response shift will also be considered during the analysis portion of this study. Oort's SEM approach, as suggested by Sajobi et al, will be conducted as a method of evaluating response shift.[22] This will be conducted following a 4-step process, as highlighted by Oort.[50] The methodology explained in his original publication will be followed in this study.

5.3.3 Subgroup Analysis

Subgroup analysis will also be performed. Individuals will be divided into groups based on the severity of disease (mild, moderate, or severe), type of hemophilia (A or B), type of treatment schedule (prophylactic or episodic), age, and frequency of bleeds. These are distinct and important characteristics that may influence PROBE's responsiveness.

5.4 Phase 4: Delphi Panel

For this phase, data will be analyzed after each consecutive step. Study investigators will evaluate the amount of agreement between the experts. Steps will continue until consensus is reached, which for the purposes of this phase will be defined as 80% agreement.[32] Once agreement is reached, this phase will be completed.

6.0 DESIGN JUSTIFICATION

6.1 Phase 1: Delphi Panel

A Delphi panel was selected to address uncertainties regarding appropriate collection points for desired patterns in QOL. This design was selected given that this information could not be sufficiently addressed using a systematic review, and other methodology would require more time and resources, resulting in a delayed implementation of the main study. This simple design can provide the necessary results in a timeline that is suitable for the goal at hand.

The eligibility for recruitment was selected in order to ensure participants are knowledgeable in the field of hemophilia. Provided that experts are required to understand the nature of the condition, the panel can include physicians and patients alike.

Consensus for this phase will be set at 80% agreement.[32] Given that much of the information collected in this phase will be confirmed during the pilot study, this level of consensus would be sufficient.

6.2 Phase 2: Pilot Study

A pilot study will be implemented in order to determine an adequate sample size for the main trial. This design was selected over a run-in period for the additional flexibility a pilot study can provide. Given that there remain some uncertainties surrounding appropriate timeline for item collection, as well as the adequacy of the selected anchors, having a pilot would not only allow for a sample size determination, but also an opportunity to troubleshoot and address some of these concerns. This way, any issues that arise could be addressed and modified before moving to the main study. With a run-in period, the main study would either continue as is, or be terminated, removing much of the flexibility and opportunity for change.

Participants enrolled in this phase will be carried over to the main study, provided no major changes are undergone. This decision was made considering the small population of individuals living with hemophilia in Canada. Having participants continue

to the main study would help ensure an adequate sample size, which may be an issue when conducting research on populations living with rare conditions.

Other specific aspects of the study design (such as inclusion/ exclusion criteria) can be found in section 6.3, given that the pilot study design will reflect that the main study.

6.3 Phase 3: Main Study

6.3.1 Study Design

Although responsiveness can be assessed using cross-sectional methods, this study has opted to use a prospective design. Unfortunately, only between-person comparisons can be made using cross-sectional methods, and it has been shown that this method underestimates within-person change.[10] In order to accurately estimate a within-person change, a prospective approach was selected.

A randomized design for this research question would have been a) impossible, as you cannot randomize individuals to experience QOL relevant event or not, and b) it would have added no value, as the change of interest happens within patients, and not across randomization. For these reasons, a randomized design could not be undergone.

A multi-center approach was selected in order to address concerns relating to adequate recruitment of participants. Given that hemophilia is a rare condition, a single center study would likely provide insufficient numbers of potential participants. In addition, using patients from across the country will contribute to the external validity of

this study. Canadian researchers and physicians looking to use PROBE in the future can be assured that our study population is likely to be representative of the general population of patients living with hemophilia.

6.3.2 Inclusion Criteria

The inclusion criteria for this study were selected based on previous studies evaluating PROBE. The questionnaire at hand was validated using a population of individuals living with hemophilia A or B, and therefore should be applied to the same population in this study.[48] Those living with other conditions will not be eligible for this study given that PROBE has not been evaluated for that population.

Although individuals must be 10 years old or older to provide reliable responses to the PROBE questionnaire, this study will recruit individuals that are 16 years of age or older.[48] This age cut-off was selected to address ethical concerns of recruiting individuals between the ages of 10 and 16. This age group would require participant assent in addition to parental consent. Given that consent is provided online before questionnaire completion, obtaining assent may prove to be more difficult using this approach. In order to facilitate the consenting process, the study population will be limited to individuals over the age of 16.

Finally, individuals will only be eligible if they are enrolled or willing to enrol in MyCBDR. Not only does this facilitate PROBE app log in, it also ensures all study data is properly captured. Individuals can technically sign up for the PROBE app without a MyCBDR account, but if said individual switches to MyCBDR mid-study, their responses from before and after the switch cannot be linked. In order to avoid the

possibility altogether, all participants will require MyCBDR access. As an added benefit, bleeds are reported and documented in the MyCBDR app and will provide the PROBE app with a more specific and targeted notification system. This way, participants will be reminded to complete PROBE as soon as they report having a bleed. This will likely improve data completion and decrease the amount of missing data. Since this was made an inclusion criterion, study staff will be able to facilitate MyCBDR registration by accepting requests on the spot, therefore limiting the wait time that currently exists.

6.3.3 Exclusion Criteria

Exclusion criteria were limited as much as possible as a way of increasing the number of potentially eligible individuals. At minimum, participants must be able to comprehend instructions, consent, and read/understand English or French. Given PROBE app is currently available for completion in these two languages for Canadian participants, it is imperative they be able to provide accurate responses in English or French.

6.3.4 Data Collection Timeline

A large part of the study timeline depends on the occurrence of specific events. Although responsiveness is typically assessed across a period in which participants receive therapy of a known effectiveness, this approach is not feasible in this field.[12] Given the chronicity of the disease, patients undergo treatment continuously throughout their lives, so a distinct period to evaluate a change in QOL before and after treatment may not exist. Bleeds and surgical interventions were selected as a substitute given these are events that are likely to result in a change in QOL.[44]

Given that the timing of these collection points is instrumental to observing the desired trends in QOL, the Delphi approach will provide confidence in the study's timeline selection. This will help ensure the timing of the PROBE questionnaires are valid and in compliance with what has been observed in practice across Canada. This approach will identify two data collection points for each event: one in the acute setting, and one in the longer term, as highlighted above. The first timepoint aims to evaluate the situation in which individuals are likely to be experiencing the negative consequences of the event, such as pain or loss of function. The second timepoint on the other hand, aims to capture the recovery process in which QOL may begin to improve. By including both of these data points, one may then evaluate PROBE's responsiveness when QOL shifts in either direction.

Finally, there is a situation in which participants may go a considerable amount of time without experiencing any events. It is possible that these individuals will still experience changes in QOL that are not prompted by bleeds or surgery. For this reason, a 6-month timepoint has been added. Having this data collection point will allow investigators to evaluate responsiveness in this case, effectively addressing the secondary objective of this study.

6.3.5 Data Analysis

It has been suggested by the literature that the anchor and the PRO in question should be correlated.[20] For this reason, suitability of the anchors selected for this study will be assessed by calculating their correlations with the PROBE score before continuing with data analysis.

For both the primary and secondary objectives, anchor-based approaches will be favoured over distribution-based ones.[20] Using an anchor, particularly a patient-reported one, allows for the patients to indicate whether they have experienced a change and whether said change was important. Using an objective measure or a physician's opinion may not entirely capture the participants' experiences. Having patients provide this insight on their own behalf assures this misinterpretation is avoided. Distribution-based methods rely on spread statistics and make an assumption that changes in QOL have occurred in a set period.[14] This is not an ideal approach for evaluating responsiveness, although the literature has indicated it can be used to support results from the anchor-based approach. For this reason, all distribution-based methods will receive less weight in the decision-making process.[20]

Of the two anchor-based approaches, the global health rating question will also be given preference over the EQ-5D as an anchor for calculating MIDs. Although the validity and reliability of the global health rating question are unknown, it has previously been shown to be sensitive to change in both directions.[14] The suitability of the EQ-5D as an anchor remains unknown.[51] It is suggested that EQ-5D may not be responsive to change in hemophilia patients, although this has not been widely studied.[51] Given that the main purpose of an anchor is to provide an indication of change in QOL, the EQ-5D may not be the best suited in the absence of more responsiveness data. For this reason, MIDs calculated using this questionnaire will only be used as supporting evidence.

These anchors will be paired with the importance question to determine not only if a change occurred, but also if that change was meaningful. Beaton et al. have highlighted

the nuance between these concepts.[10] Many studies have focused on whether a change has occurred, but have not attached an importance to this change. In order to truly calculate a change that is both minimal and important, both of these questions should be considered.

It was also decided that the selected anchors will be evaluated using mean change scores. Data collected using an anchor can be evaluated using a variety of approaches. Commonly applied are receiver operator characteristic curves, although this method requires that the responses to the anchor be dichotomized.[19,27] Participants would be placed into "improve and unimproved" or "deteriorated or not deteriorated", which loses much of the collected information.[19] Additionally, this approach does not examine minimal changes in QOL.[19] Correlation may also be used but this unfortunately evaluates the closeness to a linear relationship and may miss a close but non-linear relationship.[19]

As for distribution-based approaches, a paired t-test may be applied. This approach is unfortunately influenced heavily by the sample size.[19] Very small changes in the individual may lead to a statistically significant change given a large enough sample size.[19,25] Guyatt's responsiveness could also be calculated, but this approach requires a known MID value, which is currently unavailable.[19] ES, SRM, and SEM are the most popular approaches to evaluating responsiveness and using them in this study may facilitate future comparison of estimates in two different instruments in the same population.

6.3.6 Subgroup Analysis

It is known that MIDs are heavily influenced by the population at hand, and a single MID may not exist for a given questionnaire.[17,20,31] For this reason, it is vital to evaluate a variety of important subgroups to assess the stability of the overall MID, or to identify different MIDs for different subgroups, if needed. The subgroups that will be evaluated in this study have been selected based on their importance in the field of hemophilia.

It has been previously noted that individuals living with varying severities of hemophilia may experience events differently. In fact, those living with less severe disease may demonstrate a greater impact on QOL following an event.[52,53] It is possible that those living with more severe disease have become more proficient at managing events and may therefore not experience as much distress when one occurs. Given their ability to manage events well, individuals living with more severe disease may require more severe events in order to perceive it as important.[52,53] This may translate to a larger MID value, where a larger change in PROBE score may be required to indicate a minimally important change in QOL. For this reason, individuals living with different severities will be evaluated separately.

The literature has suggested that the clinical presentation of hemophilia A and hemophilia B may not be the same.[39–41] Given that these two groups may experience differing frequencies and/or severities of events, it is possible that the MID will not be the same for both groups. Individuals that experience events more frequently may be more experienced with their management, possibly influencing the MID. On the flip side,

frequent bleeders may be experiencing more rapid degeneration of joints than their counterparts, which may therefore influence the respective MID. For this reason, hemophilia A and B will be evaluated separately. Similarly, bleed frequency will also be considered as an important subgroup. The relationship between hemophilia type, bleed frequency, and the MID would be an important aspect to further investigate.

The treatment schedule that is used may also demonstrate a similar trend. Research has demonstrated that prophylactic treatment can maintain joint health.[46] For those undergoing prophylactic treatment, a more severe event may be required to result in a minimally important change. Given the possible influence on the MID, prophylactic and episodic schedules will be considered separately.

Age will also be considered when evaluating subgroups. With the passing years, individuals living with hemophilia experience an ongoing deterioration of the joints, with damage acting cumulatively over time. For this reason, it is possible that a bleed may not affect pain level, range of motion, or QOL of a younger individual to the same extent as an older individual. For this reason, it is possible that the MID may differ depending on the participant's age.

Subgroups were selected in order to investigate any possible factors that may influence the overall MID value. The resulting MIDs may or may not be helpful in practice. Having an overall MID value for the PROBE questionnaire would provide a quick and easy way of interpreting changes in score. Having a separate MID value for each of the subgroups could provide more specific and targeted approach to interpreting

changes in score. Having multiple MID value could, however, complicate interpretation. Selecting the most appropriate MID may prove to be difficult and selecting the most applicable subgroup may be confusing. Further, these subgroups will inevitability include small sample sizes, decreasing the confidence in the associated MID values. Nevertheless, exploring these relationships may highlight whether or not there are certain relationships that require further consideration.

6.3.7 Anticipated Results and Interpretation

For the anchor-based approaches used in both primary and secondary analysis, we anticipate that participants who indicate an improvement in QOL on the global health question or an increase in EQ-5D score will also demonstrate an increase in PROBE score. For those that indicate a deterioration in QOL, a decrease in PROBE score is also anticipated. The resulting MIDs will provide an indication of the smallest score that represents a meaningful change in QOL. The MIDs will help understand how much of a change in score indicates a true change. If an individual has a change in score that is smaller than the MID, it is likely that their QOL has not changed in that time. Only changes in score above the MID indicate a potential meaningful change.[14]

For the distribution-based approaches used in both primary and secondary analysis, MDCs will be calculated. It is anticipated that these values will be lower than the calculated MIDs. Given that these values focus on finding statistically significant changes rather than clinically significant changes, the resulting scores are likely to be lower.

In terms of interpreting the results of these approaches, the literature has consistently quoted 0.2, 0.5, and 0.8 to represent small, moderate, and large responsiveness, respectively, for the ES and SRM approaches.[14,17,19] These cut points will be used for the purposes of this study as well. As for the SEM, it has been suggested that 1 SEM approximates the MID, and this logic will also be followed in this study as well.[14,29,54]

Regression analysis will be performed in an exploratory manner. Where MIDs indicate the amount of change required in the PROBE score to see a meaningful change, regression indicated the amount of change in the anchor is associated with a one-unit change in PROBE score.[19] The magnitude of this value will aid in its interpretation, given that a value close to zero will indicate that large changes in PROBE would be required to observe any change in the anchor.[19] Following this logic, a value closer to 1 could indicate a more responsive tool.[19] It is possible that the results from these calculations will not provide any novel information, but they may help with interpretation and understanding of responsiveness.

It has been proposed that response shift can lead to under or over-estimation of PRO scores collected longitudinally. Given that there is a possibility that response shift could influence the responses to the PROBE questionnaire, it will be evaluated in this study.[22] The procedures followed in this study will help identify if response shift occurred and will facilitate the calculation of a true change. This will allow a better understanding of whether participants experienced changes in their standards for evaluating their own quality of life, their values, or their internal understanding of the concepts at hand. This in

turn, may help explain possible unanticipated results or trends, especially in terms of magnitude of the effects calculated in this study.

6.4 Phase 4: Delphi Study

A Delphi panel is being considered following suggestion from the literature published in this field.[20,31] Given the nature of the objective of this phase, this design provides an opportunity for gaining expert advice if any uncertainties arise. Given that these answers could not be gained by searching existing literature, expert consensus can provide the necessary results in a timeline that is suitable for the goal at hand.

The eligibility for recruitment was selected in order to ensure participants are knowledgeable in the PROBE questionnaire and its scoring. Provided that experts are required to understand PROBE, the panel can include physicians and patients alike.

7.0 POSSIBLEHARMS

7.1 Phase 1: Delphi Study

This phase of the study does not have any anticipated harms to participants. Expert responses will be anonymized between rounds, ensuring the confidentiality of the participants. There is a possible concern that the consensus reached during the Delphi panel may not be validated before implementation. To address this concern, the planned pilot study will act as a method to confirm the proposed timeline.

7.2 Phase 2: Pilot Study

The harms that may possibly be experienced by pilot study participants match those explained in section 7.3 (Main Study).

7.3 Phase 3: Main Study

This study does not have any anticipated direct harms to study participants. Individuals will be asked to dedicate their time to completing the questionnaires which may be fatiguing or frustrating if required frequently. The magnitude of this burden will depend on the events experienced, with those experiencing more bleeds or multiple surgeries requiring more time and energy to complete all data points.

Theoretically, there maybe a concern that information collected on the app may be accessed by a third party. Although this is unlikely, information that is collected will be anonymized and will not be connected to any identifiable information. This study will comply with all ethics rules in order to maintain participant privacy.

Finally, it is possible that participants may be taking note of their responses to the questionnaires as a method of tracking their progress or health status. This may result in some emotional distress, in the case that the individual's scores do not adequately reflect what they believe they are experiencing (eg. PROBE score worsening despite feeling they have made positive progress). For these participants, it may be a frustrating experience to track scores without the ability to properly interpret them.

7.4 Phase 4: Delphi Study

Similarly to section 7.1, this phase of the study does not include any risks to the participants. Responses to iterative stages will be anonymized in order to maintain confidentiality of the panel experts.

8.0 POSSIBLE BENEFITS

8.1 Phase 1: Delphi Study

Experts participating in this phase may not experience any personal benefits to participating in this study. They will, however, be contributing to their knowledge to further research in the field of hemophilia. Their consensus alone will provide interesting insight on the trends in QOL patients may be experiencing. Further, this information will aid in identifying whether PROBE could be used to adequately assess changes in QoL over time.

8.2 Phase 2: Pilot Study

The benefits that may possibly be experienced by pilot study participants match those explained in section 8.3 (Main Study).

8.3 Phase 3: Main Study

Individuals participating in this study may or may not experience any direct benefits. It is possible that completing the PROBE questionnaire at multiple timepoints

may help participants be more mindful of their own health status and consider areas that may require improvement. This may in turn lead to more focused and directed appointments with their physicians.

As for societal benefits, the information collected from this study will provide a better understanding of the responsiveness of the PROBE questionnaire. This information will provide more context and increase the interpretability of changes observed in the PROBE score.[17] Understanding this property will direct future uses of the questionnaire, be it in the clinical or research setting. Knowing the MID can provide a method of interpreting results from research trials evaluating the effectiveness of novel therapeutic drugs and could help physicians better understand changes in QOL at consecutive appointments.

8.4 Phase 4: Delphi Study

Similarly to section 8.1, this phase of the study does not include any benefits to the participants. The consensus that is determined in this phase will contribute to a societal benefit namely, in that changes in PROBE score can be adequately understood. Determining an MID value would prove to be beneficial to patients and physicians alike, as it would provide a method of interpreting changes in PROBE score. This knowledge could expand the use of this questionnaire as it could be applied in future research projects aimed to determine drug effectiveness, as well as in the clinical setting to track patient progress.

9.0 ETHICAL CONSIDERATIONS

Ethics approval will be obtained at each participating site before implementing this study. This process will begin at McMaster University, and once received locally, approval will be sought at all other participating CBDR sites across Canada. Recruitment will not begin at any site until approval is obtained. Informed consent will be received from every participant before completing the first PROBE questionnaire.

As with any electronic database, concerns surrounding breaches of confidentiality are often present. In theory, it is not outside the realm of possibility that an unwanted third party could gain access to data stored in such a database. That being said, individuals that respond to the PROBE questionnaire in this study do so through their MyCBDR accounts to limit this concern. In this study, PROBE questionnaire responses will be housed in one database while baseline demographics and event information will be housed in the CBDR database. The two will be linked with a unique ID code. Given that CBDR is a secure database and the PROBE database will not hold any identifying information, the concern in breach of confidentiality should be minimal. All data will also be kept on password protected computers with redundant hardware, mirrored hard disks, and secure backup systems. Backup media will be managed following standard safety procedures. Data collected on the PROBE website will be encrypted during transfer over the internet with standard methods used for data protection (https protocol).

10.0 FEASIBILITY

Certain aspects of feasibility have been mentioned throughout this proposal. First of all, PROBE has been shown to be of low burden to individuals, as completion typically requires only 15 minutes.[48] In addition, the baseline characteristics section will be removed from subsequent questionnaires to further lessen the burden of completion on individuals.

As for the integration of PROBE and CBDR, both databases share a common primary investigator and are approved by the Hamilton Health Sciences integrated REB. By appropriately extending the REB approval, integrating the PROBE questionnaire and the CBDR database should be a relatively simple task. In addition, both registries are hosted at the Health Information Research Unit at McMaster University. Integrating the two databases would require close collaboration with the IT group responsible for maintaining the CBDR database, which is located on site. This team is in the process of developing a mobile application to support longitudinal collection of PROBE, which will also contain the ability to link these registries. This application will also facilitate PROBE completion and will act as a reminder for participants.

Finally, this project also has the support of the CEO of the Canadian Hemophilia Society, a member of the PROBE steering group, and an ambassador of CBDR, which would further facilitate this process.

11.0 LIMITATIONS

It has been suggested that using patient-reported anchors to evaluate a PRO measures concurrent validity rather than responsiveness.[25] Unfortunately, using an objective anchor is not feasible for this study. In hemophilia, there are a few objective outcomes that may be evaluated, but unfortunately, they only cover one isolated aspect of the disorder. For example, using a target joint score as a method of determining which patients have improved or deteriorated would demonstrate a potential change in QOL, but only evaluates joint function and does not consider number of days of work or school missed, for example. Having physicians collect a variety of these outcomes and make a subjective decision on whether the patient has experienced a meaningful change would not only be impractical but would also be subject to inconsistencies between raters. Given this limitation, a patient-reported anchor is the best substitute available in for this study. We anticipate that using multiple anchors as well as multiple distribution-based approaches will provide a confident estimate of PROBE's responsiveness.

Obtaining an adequate sample size may also be an area of concern. Although individuals are likely to experience 2-4 bleeds a year, it is unknown what proportion of those bleeds will result in a meaningful change. Ideally, each individual would experience at least one meaningful bleed in the data collection period. Nevertheless, it is possible that these types of bleeds are less common, resulting in a higher number of participants needed to recruit. Unfortunately, if meaningful bleeds are more common, with the same individual experiencing more than one in the study period, this will not decrease the sample size required. In order to ensure independence of the data points, only one bleed

will be used per participant during data analysis. Further, it is possible that the groups chosen for subgroup analysis (namely severity of disease, type of hemophilia, age, type of treatment schedule) will not be sufficiently large to provide good estimates. This is an unfortunate limitation of working with rare conditions, such as hemophilia. Depending on the final calculated sample size, it is possible that this study will not be feasible to undertake. If the number of participants to recruit is exceedingly large, the study may prove to be too expensive to fund.

Even with the estimates gained in the planned pilot trial, it is possible the recruited participants will not experience enough events that result in important change. In addition, it is possible that there are no meaningful changes in QOL observed in the periods where no events were experienced. This is a potential limitation that is out of the investigator's hands. In order to limit the possibility of this issue occurring, the number of recruited individuals should exceed the calculated sample size.

If the number of events required is not reached in this study, power will be recalculated. The information gained from this study will be used to design a more appropriately sized international study.

12.0 STUDY IMPLICATIONS

This study will primarily help with the interpretability and clinical use of the PROBE questionnaire. Understanding the MID will allow for the implementation of this questionnaire in research trials and clinical use across the country. Knowing PROBE is responsive to change in QOL opens the field to having a PRO measure that can be used in studies of novel drug effectiveness. Physicians may also be able to implement PROBE as a regularly collected item in their clinics. This would provide a method of understanding the ways in which the patient has changed since their last appointment in an attempt to provide more specific guidance or intervention.

13.0 REFERENCES

1. Deshpande PR, Rajan S, Sudeepthi BL, Abdul Nazir CP. Patient-reported outcomes: A new era in clinical research. *Perspect Clin Res*. 2011;2(4):137-144. doi:10.4103/2229-3485.86879

2. Bottomley A, Jones D, Claassens L. Patient-reported outcomes: Assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency. *Eur J Cancer*. 2009;45(3):347-353. doi:10.1016/j.ejca.2008.09.032

3. McLeod LD, Coon CD, Martin SA, Fehnel SE, Hays RD. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11(2):163-169. doi:10.1586/erp.11.12

4. Weldring T, Smith S. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Heal Serv Insights*. 2013:61. doi:10.4137/HSI.S11093

5. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: Draft guidance. *Health Qual Life Outcomes*. 2006;4:1-20. doi:10.1186/1477-7525-4-79

6. Bolarinwa O. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Niger Postgrad Med J*. 2015;22(4):195. doi:10.4103/1117-1936.173959

7. Mokkink LB, Terwee CB, Patrick DL, et al. COSMIN definitions of domains, measurement properties, and aspects of measurement properties Term Definition Domain. :1.

8. Kline P. *The Handbook of Psychological Testing*. 2nd ed.; 2000.

9. Hankins M. Questionnaire discrimination: (re)-introducing coefficient δ. *BMC Med Res Methodol*. 2007;7:3-7. doi:10.1186/1471-2288-7-19

10. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. *J Clin Epidemiol*. 2001;54(12):1204-1217. doi:10.1016/S0895-4356(01)00407-3

11. Polit DF. Assessing measurement in health: Beyond reliability and validity. *Int J Nurs Stud*. 2015;52(11):1746-1753. doi:10.1016/j.ijnurstu.2015.07.002

12. Guyatt G, Walter S, Norman G. Measuring Change Over Time- Aseessing the Usefulness of Evaluative Instruments. *J Chronic Dis*. 1987;40(2):171-178.

13. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. *CMAJ*. 1986;134(8):889-895.

14. Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*. 2007;7(5):541-546. doi:10.1016/j.spinee.2007.01.008

15. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes*. 2006;4(Mic):3-7. doi:10.1186/1477-7525-4-54

16. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes*. 2006;4:1-5. doi:10.1186/1477-7525-4-70

17. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life—a systematic review. *J Clin Epidemiol*. 2017;89:188-198. doi:10.1016/j.jclinepi.2017.06.009

18. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. *Control Clin Trials*. 1989;10(4):407-415. doi:10.1016/0197-2456(89)90005-6

19. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol*. 2000;53:459-468. doi:10.1016/S0895-4356(99)00206-1

20. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining

responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102-109. doi:10.1016/j.jclinepi.2007.03.012

21. Schwartz CE, Bode R, Repucci N, Becker J, Sprangers MAG, Fayers PM. The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Qual Life Res*. 2006;15(9):1533-1550. doi:10.1007/s11136-006-0025-9

22. Sajobi TT, Brahmbatt R, Lix LM, Zumbo BD, Sawatzky R. Scoping review of response shift methods: current reporting practices and recommendations. *Qual Life Res*. 2017;0(0):1-14. doi:10.1007/s11136-017-1751-x

23. Machuca C, Vettore M V., Krasuska M, Baker SR, Robinson PG. Using classification and regression tree modelling to investigate response shift patterns in dentine hypersensitivity. *BMC Med Res Methodol*. 2017;17(1):120. doi:10.1186/s12874-017-0396-3

24. Schwartz CE, Andresen EM, Nosek MA, Krahn GL. Response Shift Theory: Important Implications for Measuring Quality of Life in People With Disability. *Arch Phys Med Rehabil*. 2007;88(4):529-536. doi:10.1016/j.apmr.2006.12.032

25. Gatchel RJ, Mayer TG. Testing minimal clinically important difference: consensus or conundrum? *Spine J*. 2010;10(4):321-327. doi:10.1016/j.spinee.2009.10.015

26. Gatchel RJ, Mayer TG, Choi Y, Chou R. Validation of a consensus-based minimal clinically important difference (MCID) threshold using an objective functional external anchor. *Spine J*. 2013;13(8):889-893. doi:10.1016/j.spinee.2013.02.015

27. Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. *Phys Ther*. 1996;76(10):1109-1123. http://www.ncbi.nlm.nih.gov/pubmed/8863764.

28. Zhang Y, Zhang S, Thabane L, Furukawa TA, Johnston BC, Guyatt GH. Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. *J Clin Epidemiol*. 2015;68(8):888-894. doi:10.1016/j.jclinepi.2015.02.017

29. Turner D, Schünemann HJ, Griffith LE, et al. The minimal detectable change

cannot reliably replace the minimal important difference. *J Clin Epidemiol*. 2010;63(1):28-36. doi:10.1016/j.jclinepi.2009.01.024

30. de Vet HCW, Terwee CB. The minimal detectable change should not replace the minimal important difference. *J Clin Epidemiol*. 2010;63(7):804-805. doi:10.1016/j.jclinepi.2009.12.015

31. Engel L, Beaton DE, Touma Z. Minimal Clinically Important Difference. A Review of Outcome Measure Score Interpretation. *Rheum Dis Clin North Am*. 2018:1-12. doi:10.1016/j.rdc.2018.01.011

32. Jorm AF. Using the Delphi expert consensus method in mental health research. *Aust N Z J Psychiatry*. 2015;49(10):887-897. doi:10.1177/0004867415600891

33. Curtis JR, Yang S, Chen L, et al. Determining the Minimally Important Difference in the Clinical Disease Activity Index for Improvement and Worsening in Early Rheumatoid Arthritis Patients. *Arthritis Care Res*. 2015;67(10):1345-1353. doi:10.1002/acr.22606

34. Limperg PF, Terwee CB, Young NL, et al. Health-related quality of life questionnaires in individuals with haemophilia: a systematic review of their measurement properties. *Haemophilia*. 2017;23(4):497-510. doi:10.1111/hae.13197

35. Boehlen F, Graf L, Berntorp E. Outcome measures in haemophilia: a systematic review. *Eur J Haematol*. 2014;93:2-15. doi:10.1111/ejh.12369

36. White GCI, Rosendaal F, Aledort LM, et al. Factor VIII and Factor IX of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. *J Thromb Haemost*. 2001;85(3):560. http://c.ymcdn.com/sites/www.isth.org/resource/group/d4a6f49a-f4ec-450f-9e0f-7be9f0c2ab2e/official_communications/fviiipharmaco.pdf.

37. Auerswald G, Dolan G, Duffy A, et al. Pain and pain management in haemophilia. *Blood Coagul Fibrinolysis*. 2016;27(8):845-854. doi:10.1097/MBC.0000000000000571

38. Szende A, Schramm W, Flood E, et al. Health-related quality of life assessment in adult haemophilia patients: a systematic review and evaluation of instruments. *Haemophilia*. 2003;9(6):678-687. doi:823 [pii]

39. Nagel K, Walker I, Decker K, Chan AKC, Pai MK. Comparing bleed frequency and factor concentrate use between haemophilia A and B patients. *Haemophilia*. 2011;17(6):872-874. doi:10.1111/j.1365-2516.2011.02506.x

40. Santagostino E, Fasulo MR. Hemophilia A and hemophilia B: Different types of diseases? *Semin Thromb Hemost*. 2013;39(7):697-701. doi:10.1055/s-0033-1353996

41. Tagariello G, Iorio A, Santagostino E, et al. Comparison of the rates of joint arthroplasty in patients with severe factor VIII and IX deficiency : an index of different clinical severity of the 2 coagulation disorders. 2009;114(4):779-784. doi:10.1182/blood-2009-01-195313.An

42. George L, Sullivan S, Giermasz A, et al. Hemophilia B Gene Therapy with a High-Specific-Activity Factor IX Variant. *N Engl J Med*. 2018;377(23):2215-2227. doi:10.1056/NEJMoa1708538

43. Rangarajan S, Walsh L, Lester W, et al. AAV5–Factor VIII Gene Transfer in Severe Hemophilia A. *N Engl J Med*. 2018;377(26):2519-2530. doi:10.1056/NEJMoa1708483

44. Neufeld EJ, Recht M, Sabio H, et al. Effect of Acute Bleeding on Daily Quality of Life Assessments in Patients with Congenital Hemophilia with Inhibitors and Their Families: Observations from the Dosing Observational Study in Hemophilia. *Value Heal*. 2012;15(6):916-925. doi:10.1016/j.jval.2012.05.005

45. Rocha P, Carvalho M, Lopes M, Araújo F. Costs and utilization of treatment in patients with hemophilia. *BMC Health Serv Res*. 2015:1-7. doi:10.1186/s12913-015-1134-3

46. Manco-Johnson MJ, Abshire TC, Shapiro AD, et al. Prophylaxis versus Episodic Treatment to Prevent Joint Disease in Boys with Severe Hemophilia. *N Engl J Med*. 2018;357(6):535-544.

47. Pocoski J, Benjamin K, Michaels LA, Flood E, Sasane R. An overview of current trends and gaps in patient-reported outcome measures used in haemophilia. *Eur J Haematol*. 2014;93(S75):1-8. doi:10.1111/ejh.12323

48. Skinner MW, Chai-Adisaksopha C, Curtis R, et al. The Patient Reported Outcomes , Burdens and Experiences (PROBE) Project : development and evaluation of a questionnaire assessing patient reported outcomes in people with haemophilia. *Pilot Feasibility Stud*. 2018;4(58):1-10.

49. Terwee CB. COSMIN checklist with 4-point scale. *Cosmin*. 2011:6.

50. Oort FJ. Using structural equation modeling to detect response shifts and true change in discrete variables: an application to the items of the SF-36. *Qual Life Res*. 2016;25(6):1361-1383. doi:10.1007/s11136-015-1195-0

51. Payakachat N, Ali MM, Tilford JM. Can The EQ-5D Detect Meaningful Change? A Systematic Review. *Pharmacoeconomics*. 2015;33(11):1137-1154. doi:10.1007/s40273-015-0295-6

52. Albrecht GL, Devlieger PJ. The Disability Paradox: Highly Qualified of Life against All Odds. *Soc Sci Med*. 1999;48:977-988. doi:10.1016/S0277-9536(98)00411-0

53. Roush SE, Sharby N. Disability Reconsidered: The Paradox of Physical Therapy. *Phys Ther*. 2011;91(12):1715-1727. doi:10.2522/ptj.20100389

54. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol*. 1999;52(9):861-873. doi:10.1016/S0895-4356(99)00071-2

55. Streiner DL, Norman GR. Measuring change. In: *Health Measurement Scales*. Oxford University Press; 2008:277-298. doi:10.1093/acprof:oso/9780199231881.003.0011

APPENDIX A- DISTRIBUTION-BASED FORMULAS[55]

SEM= $SD_{(Baseline)}$ x $\sqrt{1-R}$

ES= mean difference in scores / $SD_{(Baseline)}$

SRM= mean difference in scores/ $\sqrt{2(1-R)}$

APPENDIX B- MEASUREMENT PROPERTIES OF HEMOPHILIA QUESTIONNAIRES[34]

Table 5 Measurement properties of hemophilia questionnaires previously evaluated in the literature

Questionnaire	Internal Consistency (total score)	Measurement Error	Test-Retest Reliability	Content Validity	Structural Validity	Hypotheses Testing	Cross-Cultural Validity	Responsiveness
CHO-KLAT	?	N/A	+++	+++	N/A	+++	N/A	N/A
Haemo-QoL I	+	N/A	N/A	+++	?	+	N/A	N/A
Haemo-QoL II	+	N/A	+	+++	?	+	N/A	N/A
Haemo-QoL III	+	N/A	+	+++	?	+	N/A	N/A
Haemo-QoL Index	?	N/A	+	N/A	---	+	N/A	N/A
Hemofilia-QoL	+++	N/A	+	+++	N/A	+	N/A	N/A
Hemophilia Well-Being Index	+++	N/A	++	+++	+++	+	N/A	N/A
HAEMO-QoL-A	+++	N/A	++	+++	++	+	++	N/A
Haem-A-QoL	++	N/A	+	N/A	N/A	+	N/A	N/A
Haem-A-QoL (elderly)	++	N/A	++	+++	N/A	+	N/A	N/A

+++ or --- = strong evidence positive/ negative result, ++ or -- = moderate evidence positive/ negative result, + or - = limited evidence positive/ negative result, ? = unknown due to poor methodological quality, N/A = no information available