A SYMPTOM-BASED OUTCOME MEASURE FOR CLINICAL TRIALS IN NONULCER DYSPEPSIA

DONALD GARTH MACINTOSH

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

A SYMPTOM-BASED OUTCOME MEASURE FOR CLINICAL TRIALS IN NONULCER DYSPEPSIA.

BY

© DONALD GARTH MACINTOSH MD.

A thesis submitted to the School of Graduate

Studies in partial fulfillment of the

requirements for the degree of

Master of Science.

Department of Community Medicine

Memorial University of Newfoundland

October 1997

St. John's Newfoundland

Abstract:

Purpose: The purpose of this study was to develop and validate a symptom-based outcome measure for clinical trials in nonulcer dyspepsia (NUD).

Methods: Patients referred to the GI outpatient clinic with chronic upper abdominal pain were approached for this study. Participants with normal endoscopy who met the inclusion criteria were enrolled and classified into standard subgroups of NUD. The ulcerlike subgroup received acid suppressive therapy and the dysmotilitylike subgroup received prokinetic agents.

Each subject completed a questionnaire at the initial visit (T1), one week later before treatment (T2), and after one month of treatment (T3). Each subject selected the symptom most important to them. The frequency and severity of the selected symptom was recorded as was a global assessment by the subject of their overall status. Other data recorded included a physician global assessment, the subject's and physician's impression of change in symptoms with treatment, and the subject's antacid use.

Results: Forty-four subjects were enrolled. The primary outcome measure of the study was the product of the selected symptom's frequency and severity. Instrument reliability was assessed by Spearman rank correlation r=0.85 and the intraclass correlation coefficient = 0.83.

Good correlation between the measure and patient global assessment (|r|=.596) and between the measure and patient assessment of treatment response (|r|=.584) was noted. The physician global assessment was moderately related to the measure (|r|=.437). Some relationship was seen between the measure and physician assessment of treatment response (|r|=.329). There was no relationship with change in antacid use (|r|=.143).

Conclusion: The main measure was reliable in untreated subjects and responsive in subjects who responded to therapy. This measure also appeared to be valid. The method of combining frequency and severity of the subject-selected symptom into a product was shown to be a useful means of assessing treatment effect in NUD patients. This study documents a reliable, responsive, and valid new outcome measure for use in clinical trials of NUD.

Acknowledgments:

I would like to acknowledge the help I have received from a number of individuals during this project. In particular, I would like to acknowlege the guidance, support, and encouragement received from my supervisor, colleague, and friend, Dr. John Fardy. This work would not be what is today without his advice. I wish to thank the members of my supervisory committee, Dr. Brendan Barrett and Dr. Ian Bowmer, for their recommendations and helpful hints. I wish to acknowledge the financial support received from the General Hospital Foundation which allowed me to start this project. Finally, I wish to thank my wife, Dr. Rebecca MacIntosh. Rebecca's love, understanding, and encouragement gave me the purpose to finish this work. This is very much hers.

Table of Contents:

List of Tables and Figures:

TABLES

FIGURES

List of Abbreviations:

NUD	Nonulcer Dyspepsia
SIP	Sickness Impact Profile
QALY	Quality-adjusted Life-year
VAS	Visual Analog Scale
IBD	Inflammatory Bowel Disease
IBDQ	Inflammatory Bowel Disease Questionnaire
IBS	Irritable Bowel Syndrome
GERD	Gastroesophageal Reflux Disease
ICC	Intraclass Correlation Coefficient
NSAID	Nonsteroidal Anti-inflammatory Drug
DIBS	Duration Intensity Behaviour Scale
T1	Time 1. At the initial visit and assessment
T2	Time 2. After one week and before onset of therapy
T3	Time 3. Final visit after one month of therapy

Chapter 1 INTRODUCTION:

1.1: Measurement of Health:

Medical research today often involves the administration of an intervention to a predefined group of individuals followed by the measurement of response to that intervention. The desired outcome is an improvement in some aspect of the trial participants' health. To determine the magnitude of an observed response or lack of response requires measurement of a health outcome.

A variety of outcomes have been used in clinical trials including laboratory tests, objective clinical measures such as morbidity and mortality rates, and more recently, subjective measures of symptoms and quality of life. Outcome measurement in the laboratory supplies objective measurements, for example hemoglobin levels or blood glucose, which provide little inherent difficulty to the researcher. Subjective judgment is generally not required. Clinical trial research in the past concentrated on assessments which usually did not require subjective opinions. These outcomes, for example, might be prolongation of life or prevention of a life threatening condition such as cancer or heart disease. Measurement of mortality and occurrences of clearly defined conditions such as a myocardial infarction was relatively straightforward. In more recent times, the situation has become more complex. The effect of new drugs on quantity of life is likely to be marginal. Now researchers are more aware of the importance of the impact of an intervention on quality of life (1). Quality of life is now used as a primary outcome measure

to determine patient response to various interventions. Such outcomes have been utilized in a variety of trials in areas as diverse as cancer (2), joint disease (3), heart disease (4), and chronic lung disease (5). Quality of life however cannot be measured in the laboratory. New scientific methods are necessary to measure quality of life in a consistent, meaningful manner.

To begin to measure quality of life first requires an understanding of what exactly constitutes health. Most widely accepted in the past was the traditional medical model of health. This model defined health as the absence of disease and infirmity. The medical model has been criticized for ignoring social health and the importance of interactions with others (6). A recent holistic model includes physical as well as mental and social aspects of health as defined by the World Health Organization. The World Health Organization definition states health is "A state of complete physical, mental, and social well-being and not merely the absence of disease and infirmity" (7).

To measure quality of life obliges the investigator to consider all aspects of health. The circumstances of a person's health can be assessed objectively by the physician or subjectively by the person. One's subjective perception of health can be influenced by experience, belief, or expectations (8). Expectations regarding health and one's ability to cope with limitation can affect perception of health, therefore two people with the same objective health status may have very different qualities of life. Relying solely on objective data may omit relevant details such as a patient's ability to tolerate discomfort (8). Physical and psychological components of disease are

not mutually exclusive and their impact on a person's impression of health should not be ignored.

Today, in medical research, situations may confront the investigator where no objective measures of illness exist and where quality of life measures are either not available or not applicable. In this instance the investigator must measure the symptomatic response to an intervention. Functional bowel diseases provide a good example. These conditions consist of variable combinations of gastrointestinal complaints for which no cause is known. Since there are no objective measures of patient illness, no objective measures of response to an intervention are available. The researcher therefore must use subjective symptom-based outcome measures to determine the magnitude of a response to therapy.

1.2: Types of Quality of Life Measures:

Quality of life measurements fall into two categories, generic and disease-specific instruments. Generic instruments are designed to sample the complete spectrum of function or disability relevant to quality of life. Disease-specific instruments concentrate on certain conditions or factors related to the condition or disease of interest.

Health related quality of life instruments are made up of a number of items or questions. These items are grouped in a number of domains or dimensions which refer to the area of behaviour or experience being measured. Domains might include physical function or emotional function, for example (9).

There are two main types of generic instruments, health profiles and utility measures. Health profiles use a single instrument to measure different dimensions or aspects of quality of life. A common scoring system is shared which can be aggregated into a small number of scores or even a final single score or index. An example is the Sickness Impact Profile (SIP) (10). The SIP assesses sickness-related dysfunction in 12 different categories producing a score for each category. Various categories can be combined into a physical dimensions score, a psychosocial dimension score, and an overall score with independent categories of work, eating, sleep and rest, home management, and recreation and pastimes (11). Overall SIP scores for well populations are low and scores for patients with chronic debilitating illnesses such as rheumatoid arthritis are high.

The SIP has been used in studies of cardiac rehabilitation (12), total hip arthroplasty (13), and treatment of back pain (14).

Health profiles allow determination of the effects of an intervention on different aspects of quality of life such as emotional dysfunction without the need for multiple instruments. Health profiles can also be used in a wide variety of conditions (15). There are also disadvantages associated with the use of health profiles. The generality of their design limits their applicability to more specific aspects of quality of life. In addition, health profiles may not focus on specific aspects of interest, therefore resulting in an unresponsive instrument that fails to detect small but clinically important changes in quality of life (16, 17).

Utility measures, which are derived from economic theory, indicate the preferences of patients for treatment and outcome. Quality of life is measured holistically as a single number along a continuum of death 0.0 to full health 1.0. There are two approaches to utility measures. One method involves asking a number of questions and depending on their responses, patients are assigned to one of a number of categories. These categories have previously been assigned utility values by a different sample of raters. An example of a utility measure using this approach is the widely used quality of well-being scale (18). The second approach entails asking patients to assign a single utility which describes their quality of life. This can be accomplished a number of ways. One method is the standard gamble in which subjects are asked to choose between their own health state and a gamble that they may die immediately or achieve

full health for the rest of their lives. The quality of life is determined by the choices made as the probabilities of immediate death or full health are varied. A simpler method involves the time trade-off method. Subjects are asked how many years in their present health state would they be willing to trade-off for a shorter life span in full health (19).

The major advantage of utility measures is their ease of use in cost-utility analysis, in which the cost of an intervention is related to the number of quality-adjusted life-years (QALYs) gained (15). Utility measures are limited in a number of ways. The measurements can vary depending on how they were obtained, therefore the validity of a single measurement can be questioned (20). Secondly, one cannot determine which aspects of quality of life are responsible for changes in utility. Finally, utility measures may not respond to small but clinically important changes in patient status (21).

Disease-specific instruments focus on limited aspects of health status. The rationale for focusing on specific items lies in the potential of increased responsiveness to change. The instrument may concentrate on certain aspects of interest to the investigator such as a specific disease (e.g. chronic lung disease), a specific population of patients (e.g. the frail elderly), or a specific problem (e.g. pain) caused by various diseases. Disease-specific instruments have been developed for many conditions including cardiovascular disease (22), arthritis (23), and cancer (2, 24).

Disease-specific instruments have a number of advantages and have been proven useful in clinical trials (22, 25). Disease-specific instruments concentrate on certain features found in the patient group of interest. For example, in chronic lung disease a quality of life measure included details of dyspnea, day-to-day activities, fatigue, and emotional problems (25). The items used were generated from input of patients with chronic obstructive lung disease who selected questions important to them. Recipients of the measure were questioned about activities which make them short of breath and how short of breath they have been doing selected activities in the two weeks prior to being assessed.

As well, specific measures can concentrate on symptoms of interest to physicians such as bowel dysfunction in patients with inflammatory bowel disease. When the Inflammatory Bowel Disease Questionnaire was being developed, an expert panel of care givers was asked to list symptoms of IBD thought to be important to them. Consequently, because of the concentration on specific, important items, disease-specific measures demonstrate increased responsiveness to changes in quality of life (26, 27). The major disadvantage of disease-specific measures is that this type of quality of life instrument is not comprehensive. Disease-specific instruments cannot compare across different diseases and therefore have limited generalisability.

Quality of life instruments can also be categorized as discriminative, predictive, and evaluative indices depending upon their role. Discriminative indices distinguish between individuals or groups

with respect to an underlying dimension when no gold standard exists. An example would be an index applied to patients following myocardial infarction to divide them into those with good and those with poor quality of life, with a view to possible intervention in the latter group. Predictive indices classify individuals into a set of pre-defined measurement categories either at initial assessment or at some time in the future. A gold standard is usually compared to determine whether these individuals have been classified correctly. An example might be the use of intelligence testing by I.Q. measurements in order to predict future performance in university. The future performance of the participants in university would be the reference measure or gold standard. Evaluative indices measure the magnitude of any longitudinal change in individuals or groups. Evaluative indexes are the main type of quality of life measurement used in clinical trials (21). An example would be the Inflammatory Bowel Disease Questionnaire (IBDQ) which has been developed for use in patients with Inflammatory Bowel Disease (28). The IBDQ examined four aspects of patient lives: symptoms related directly to the primary bowel disturbance such as frequency of defecation or abdominal cramping, systemic symptoms such as malaise or fatigue, emotional function, and social function. Administration of the instrument generates scores which have been shown to change as the patient's status changed. A worsening in the IBDQ scores correlates with worsening of the patients disease state.

1.3: Symptom-Based Outcome Measures:

A variety of approaches have been used to measure symptoms in clinical research. An overview of this type of measurement and its development is well described by Streiner and Norman and the following section is condensed from this work.(1)

When developing scales to measure symptoms it is necessary to consider the possible responses. Responses can be basically divided into categorical responses such as religion and marital status and continuous variables like blood pressure and hemoglobin levels. The second feature to be considered is the "level of measurement". If the response consists of named categories such as particular symptoms or job classifications, the variable is referred to as a nominal variable. If the variable consists of ordered categories, such as colon cancer staging, it is an ordinal variable. This is in contrast to variables where the interval between the response and the constant is known. In this situation the variable is referred to as an interval variable. An example is measurement of body temperature using the Celsius scale. The final type of variable is the ratio variable where there is a meaningful zero point so that a ratio of two responses has meaning (e.g. temperature Kelvin). The important distinction lies between nominal and ordinal data which are considered as frequencies in individual categories and use non-parametric statistics for analysis, and interval and ratio variables which are continuous variables and require parametric statistics for analysis.

One simple form of a categorical scale is a yes-no answer. The response would result in a nominal scale of measurement. The most common error made when using this type of measure is use of categorical questions when the response is not categorical. For example the question "Do you have trouble climbing stairs?" ignores the fact that there are varying degrees of trouble climbing stairs. The researcher probably wishes to find out how much trouble one has climbing stairs. Ignoring the continuous nature of many responses leads to two problems. Different people will have different ideas of what constitutes a positive response, therefore error is introduced. Secondly, error is introduced by the limited number of available responses, leading to loss of information and less reliability.

Three catagories of methods are available to researchers using continuous variables of interest. These include direct estimation techniques in which the subject is required to indicate their response by a mark on a line or a check in the box, comparative methods in which subjects choose among a series of alternatives which have been previously calibrated by a separate criterion group, and econometric methods in which numerical values are assigned to various health states to allow determination of cost/benefit ratios. The direct estimation techniques have been used widely by clinical researchers for symptom-based outcome measures. These methods include visual analog scales (VAS) and adjectival scales.

Visual analog scales consist of a line of fixed length, usually 10 cm., with anchors such as "no pain" and "pain as bad as it could be" at the extreme ends with no words describing the intermediate portions.

(See figure 1 below.) The respondents are required to place a mark on the line corresponding to their perceived state. Visual analog scales have been used in medicine to assess a variety of constructs such as pain (29), mood (30), and functional capacity (31).

Figure 1.1: Example of visual analog scale.

Visual analog scales probably provide no more valid information than well designed adjectival scales, however they are very popular because their design and use is relatively simple. Unfortunately, not all patients find VAS simple. Huskisson reported 7% could not complete a VAS rating pain severity although no details were given as to the reasons why (29). A change of categories on a discrete scale may be easier to grasp intuitively then a change of 10 to 20 mm. on a 100 mm. line of the visual analog scale.

Adjectival scales consist of adjectival descriptions and discrete or continuous responses. These specific scales are very popular and are in widespread use. Adjectival scales are measures where the rater expresses an opinion by rating his agreement with a series of statements (See figure 2). Adjectival scales are widely referred to as

Likert scales. The use of the term "Likert scale" in lieu of categorical or adjectival scale is incorrect. A Likert scale is one type of adjectival scale which uses responses framed as a continuum of agree to disagree (32). The example of an adjectival scale seen below in Figure 2 is a Likert scale. In this example, the given responses range from agree to disagree. This is contrast, for example, to a scale with the response options of: none, mild, moderate, severe, and very severe.

Figure 1.2: Example of an adjectival scale.

In your opinion, is Jean Chretien worse than previous prime ministers of Canada?

☐	☐	☐	☐	☐
strongly agree	agree	neutral	disagree	strongly disagree

Adjectival scales are popular tools for measuring symptom change in clinical trials. They are easy to design, are easily understood by subjects, and require less pre-testing than comparative methods.

1.4: Characteristics of a Subjective Health Measure:

Whenever subjective endpoints such as symptoms or quality of life are used to determine outcome, three requirements must be met by the measure (33, 34). The measure must be reproducible. It should be responsive or able to detect change. Finally, the measure should be valid i.e. measure what it is supposed to measure.

Reproducibility or reliability can be defined as the extent to which a measuring procedure yields the same results on independent repeated trials under the same conditions (11). While repeated measurements of the same phenomena are never exactly the same, they should be consistent. The difference that arises or variance can be explained by error. Any measurement will contain a certain amount of chance or random error. The amount of random error is inversely related to the reliability of the instrument. For example, if a rifle fires shots widely scattered around the target, it is considered to be unreliable. If, however, the shots are concentrated around the target, the rifle is considered to be reliable. Random error is also unsystematic, therefore the rifle shots will not deviate in a systematic or consistent fashion. Another type of error is non random or systematic error. For example, the rifle whose shots are aimed at a bull's eye, but are clustered together three inches away from the target, is affected by some type of non random error or bias.

Classical measurement theory presupposes that an observed score consists of the true score plus an error term. This leads to the

reliability coefficient as the ratio of the true variance to the true variance plus the error variance. It is too simple however to assume that all variance in scores can be neatly divided into true and error variance. The generalizability theory of Cronbach (35) states that in any measurement there are multiple sources of variance. A goal of measurement is to identify and measure these various components in order to implement strategies to reduce their effects on the measurement (1). This ultimately leads to a more reliable measure.

Responsiveness, or sensitivity to change, refers to the instrument's ability to detect clinically important change over time or after treatment. Responsiveness is determined by two properties. To be responsive an instrument must be reliable. Secondly, an instrument must register changes in score when a subject's condition changes for better or worse (36). The responsiveness of an instrument may be limited by ceiling or floor effects. For example, a ceiling effect might occur if patients with the best score still have substantial impairment. Further improvement would not be reflected by an increased score. Likewise, floor effects can occur when patients who already have the worst possible scores deteriorate further (9). In other words, the instrument must also be able to detect the full range of possible responses to be truly responsive.

Validity is a necessary property of any test or instrument. The instrument must measure what it is supposed to measure (37). There are several types of validity which are relevant for subjective measures.

Criterion validity refers to the correlation of a new measure with some other reference generally accepted as the best available measure of the disorder under study. Criterion validity can be divided into two types, concurrent and predictive. Concurrent validity involves correlation of the new measure and the criterion at the same time. For example, a new scale of depression could be administered simultaneously with the Beck Depression Inventory (an accepted measure of depression). Predictive validity infers that the criterion will not be available until sometime in the future, as one would see with a new diagnostic test where it is necessary to wait for a post-mortem examination to confirm predictions (1). Criterion validity has been used mainly to analyze the validity of certain types of tests and selection procedures. Unfortunately, in many instances in clinical medicine, a relevant criterion does not exist.

A second type of validity is content validity. Content validity examines the extent to which the domain of interest is sampled by the instrument, including choice of and importance of items on a questionnaire (21). For example, a test of arithmetic operations would not be content valid if only addition problems were assessed, excluding problems of subtraction, division, and multiplication. A measure that includes a representative sample of content domains lends itself to more accurate inferences. If important aspects are missing, the researcher is more likely to make some inferences that are wrong (1).

Face validity refers to whether items on a questionnaire appear to make sense and can be easily understood (38). Is the questionnaire

easy to use? Are any of the questions or items in the measure confusing or unclear? These questions can be answered with a small pilot study before further testing occurs.

When criterion validity is not applicable, the most rigorous means of establishing validity is construct validity. A construct is a theoretically derived notion of the dimension to be measured. Determining construct validity involves examining the relationship that might exist between the instrument and the patient group to be studied. The investigator hypothesizes how the instrument should relate to other measures. These measures or constructs are applied to the population of interest. Validity is strengthened or weakened when the hypotheses are confirmed or refuted. To demonstrate construct validity of an new instrument for patients with heart failure, an investigator may want to show that patients with poorer exercise testing score lower in aspects of the new measure that relate to physical function, and that global ratings of quality of life by the patient, relatives, and health workers bear a close relation to the results of the new index (21). The validity of an evaluative instrument is suggested when changes in the instrument correlate with changes in the other related measures (9).

Validity is not an all or nothing situation. Validation continues with further use of the instrument as in future clinical trials. The more an instrument is used and the more situations in which it is used, the greater the confidence in its validity. Guyatt has stated "perhaps we should never conclude that a questionnaire has "been validated" but

rather we should suggest that strong evidence for validity has been obtained in a number of different settings and studies" (9).

Chapter 2: DEVELOPMENT OF A SUBJECTIVE HEALTH MEASURE:

2.1: Introduction:

The actual technique of developing a subjective instrument for use in clinical trials has been described in detail by Guyatt (26). The strategy of developing such indexes involves six stages (37): item selection, reduction of number of items, questionnaire format, pretesting, reproducibility and responsiveness, and validity. These will be discussed individually in the following sections.

2.2: Item Selection:

Items used on a questionnaire must be representative of the problems faced by patients with the condition to be studied. Interviews can be performed on a random sample of patients in order to determine items of interest. Patients are asked to rate the frequency and importance of each item to themselves using a adjectival scale (very important to not important at all). A second approach is to compile a set of items before interviewing the patient sample. Appropriate items may be obtained by reviewing the current literature, polling content experts, and/or examining pre-existing questionnaires for patients with similar conditions. An example of a measure developed using this process is the Inflammatory Bowel Disease Questionnaire. A list of items for use in this questionnaire was generated by administering questionnaires to clinicians who cared for patients with inflammatory bowel disease on a daily basis and to patients with inflammatory bowel disease, reviewing the literature of inflammatory bowel disease, and utilizing items from other disease-specific instruments (39).

2.3: Reduction of the Number of Items Selected:

Generally the item selection process yields more items than can be used. The number of items may be reduced, however certain criteria are important to maintain an appropriate sample. These include the number of subjects who selected the item (item frequency), the importance attached to each item, and the relevance of each item. An approach which has been used is to combine frequency and importance criteria by multiplying the frequency of each item by its mean importance. Items with the greatest frequency-importance product are retained for the final questionnaire.

In order to assess an intervention with specific goals, items relevant to those goals must remain in the questionnaire. Each dimension or symptom to be measured must be adequately represented to reduce the variability of response and to reduce the impact of idiosyncratic responses.

2.4: Format of the Instrument:

To ensure responsiveness, each scale must be able to detect small changes if these occur. A number of characteristics are recommended for subjective scales to improve responsiveness. Scales must be composed of individual elements which are clearly defined. In other words, the ranks must be non-overlapping and discrete. If not, the scale may lead to ambiguity (40). The scale must have sufficient range in order to encompass the spectrum of possible responses in the study population. The measure must also be able to equally detect improvement and deterioration. Each scale should consist of five to seven categories. A number of studies have shown that reliability falls as less categories are used, particularly with less than five scales. An upper limit of 10 to 15 categories has been suggested to reduce confusion and complexity (1). Finally, scales should be symmetrically designed. An asymmetric scale may bias results. For example, the following scale used in a clinical trial (excellent, good, no change, worse) provides two opportunities for improvement and only one for deterioration. This increases the likelihood of a positive response and may introduce bias (40).

2.5: Pretesting:

Pretesting a group of respondents similar to those about to be studied can identify potential problems or misunderstandings. Confusing or embarrassing questions can be detected. This process generally involves testing a random sample of a few to as many as 20 patients. Poor wording or choice of response options can lead to incomplete use of all possible responses in a question. Questions can be examined to ensure that a full range of response options are used. Any modifications required can be retested prior to use.

2.6: Reproducibility and Responsiveness:

An instrument must be proven to be reproducible (reliable) and responsive to change before use in a clinical trial. To demonstrate reliability or reproducibility, the questionnaire must be repeatedly administered to a group of patients who are deemed to be stable. The intervals and number of administrations should mirror what is planned for the clinical trial. The minimum interval between administration should be at least one week. The data from this study will yield an estimate of the variability in stable patients and hence an indication of the reliability of the instrument (26). Reliable instruments will generate similar results in stable subjects on repeated administrations of the instrument.

A second assessment is required to evaluate responsiveness. In this instance, the questionnaire is given to patients similar to those in the planned trial before and after utilization of an intervention known to be efficacious. Ideally, the instrument will show not only improvement in symptom scores, but also a sufficiently large improvement relative to the variability shown by stable patients. The ratio between the change seen in the second study (patients who change with treatment) to the variability in first study (stable patients) provides an estimate of questionnaire responsiveness. The larger the difference in instrument score in the group with real change (the signal) the greater the responsiveness of the instrument. As the difference in score in subjects who are stable (the noise)

becomes larger, the reliability of the instrument becomes lower and hence the instrument less responsive (26).

2.7: Validity:

The easiest means of demonstrating validity is to compare to an accepted reference measure or "gold standard." Unfortunately, a gold standard is usually not available. In this case, a researcher should use construct validity. Does the instrument behave in relation to other measures as one would expect? Construct validity involves comparisons between measures with examination of the relationship hypothesized to exist between the measure and its constructs (26). Since evaluative instruments measure change, the relationship of change in the instrument and change in other variables should be evaluated and compared (37). Validity of an evaluative instrument can be demonstrated by showing that changes in the instrument correlate with changes in other related measures in the predicted amount and direction.

Construct validation was used to assess validity during the development of the Inflammatory Bowel Disease Questionnaire (39). The IBDQ included 30 items directed at four domains: bowel symptoms, systemic symptoms, emotional function, and social function. The IBDQ was administered to 42 patients with IBD and repeated one month later. In addition, the investigators applied patient-based global ratings of change in function, global ratings of change by the physician and relative, a disease activity index, (41, 42) and the emotional function domain of a general quality of life measure (43).

At the time the investigation was planned, the investigators made predictions about how the IBDQ should relate to changes in the other measures if this questionnaire was measuring quality of life. An example of the predictions made was that the patient's global rating of change in disease should correlate closely (correlation coefficient > 0.5) with change in the bowel symptom dimensions of the IBDQ. Of the 10 predictions made in this study, 3 were correct. In five, correlation was slightly lower than predicted. The authors concluded the results provided moderate support for the validity of the questionnaire.

Chapter 3: STATISTICAL ISSUES:

3.1: Introduction:

There are a number of statistical issues which should be addressed in the development of a subjective instrument for use in a clinical trial. Reliability, responsiveness, and validity need to be demonstrated and the role of the measure as a discriminative or evaluative instrument should be considered. If the instrument's role is to discriminate among subjects, between subject variability is important. For an evaluative instrument where change within subjects is important, the ability of the instrument to detect change must be quantitated. Finally, issues germane to use of the instrument in a controlled clinical trial such as sample size calculation will be discussed.

3.2: Reproducibility:

Reliability, or reproducibility is determined by the extent that a measuring procedure yields the same results on independent repeated trials under the same conditions (44). Reproducibility can be measured by serial administration of a test to a group of subjects believed to be stable. For parametric data produced by most quality of life instruments, reproducibility can be quantitated by Pearson's correlation coefficient (r) (45). The Spearman rank correlation coefficient provides a similar assessment with nonparametric data (46).

Measures of internal consistency such as Cronbach's alpha (47) have been commonly used in the literature to test the reliability of a measure. These measures can be calculated from a single administration of a questionnaire without the requirement of two or more administrations. Sources of variance which occur from day to day do not enter into the calculation of these measures. Consequently, measures of internal consistency should not be used to assess reliability of a subjective health measure (1).

Pearson's correlation coefficient can be used to quantify reliability, however it fails to take into account variability in results attributable to systematic, as opposed to random, differences in test scores with multiple applications. Such systematic changes can be produced by learning effects, for example. Pearson's r is also restricted to the case of two measurements per subject (36).

The intraclass correlation coefficient (ICC) which reflects both systematic and random differences in test scores is now generally accepted as the preferred method of quantitating reliability (48). The ICC can be calculated as the ratio of the variance in subject score attributable to characteristics of the subjects to the total variance in score (including variance attributable both to between subject difference and to differences among subjects) over multiple repetitions of the test. Therefore, rather than measuring the correlation between two sets of scores (as with Pearson's r), the intraclass correlation coefficient tells about concordance or the extent to which repetition of the test yields the same values under the same conditions in the same individuals. The ICC is applicable no matter how many measurements per subject, as long as there are at least two (36). For these reasons, the intraclass correlation coefficient is the recommended statistic to assess the reliability of an instrument.

3.3: Responsiveness:

To use quality of life and symptom-based assessments in a clinical trial, the researcher needs an evaluative instrument which is capable of detecting change within subjects over time. Measuring only the reliability of such an instrument is inadequate when assessing the usefulness of an instrument for this purpose. The likelihood of detecting clinically important treatment effects or the instrument's responsiveness must also be assessed.

Researchers have most commonly demonstrated responsiveness by comparing instrument scores before and after an intervention. An improvement in score would be evidence of responsiveness. A t-test has been used to detect significance. When trying to select the best instrument among several possible choices, the instrument with the largest paired t-statistic is judged the most responsive (13). However, this method does not account well for variability in scores that may occur in apparently stable patients (learning effects for example).

Another method which has been used for comparing responsiveness of competing measures is the effect size. Effect size relates changes in mean score of the instrument (from baseline measurement to measurements after the intervention) to the standard deviation of baseline scores (49). The usual calculation of the effect size takes the difference in means and divides it by the standard deviation of baseline scores. This transforms score change into a unit of measurement which could be compared with score changes of other

instruments (50). This statistic has been widely used in the social sciences.

A variant of the effect size has been suggested by Guyatt with a different denominator: the standard deviation of score changes among stable subjects. Guyatt and colleagues suggest that responsiveness is not a function of the baseline standard deviation of scores but of the variability in score changes for stable subjects (36). This approach acknowledges that there is some nonspecific variability in scores and that to be truly responsive an instrument must detect changes above and beyond this nonspecific degree of change. This responsiveness index has been referred to as the Guyatt responsiveness statistic (50).

When stable subjects not exposed to an intervention are repeatedly given the same questionnaire over time, there are inevitably changes seen in questionnaire scores. This can be due to a number of factors. A learning effect with repeated administration of a questionnaire may improve scores. In the setting of a clinical trial, improvements may be seen due to placebo and Hawthorne effects (51). The Hawthorne effect is defined as the tendency for people to change their behaviour because they are the target of special interest and attention in a study (52). The Guyatt statistic has the advantage of accounting for nonspecific score changes often noted in patients who are clinically unchanged (50).

Guyatt's variant of effect size can be calculated in the following manner. The same numerator for effect size calculations is used, that

is the difference in means measured before and after an intervention, and is divided by the standard deviation of score changes for stable patients. It should be noted that this standard deviation is equivalent to the square root of twice the mean square error. The data for calculation of the numerator should be obtained from study subjects who have demonstrated improvement from the intervention by some external criterion. Deyo suggested patient-clinician consensus to select the subgroup of improved patients for estimating responsiveness or more properly effect size (50). The data for calculating the standard deviation of score changes in stable subjects can be obtained from the subset who did not improve with the intervention. Alternatively, this information could be obtained from data used to calculate reliability statistics (untreated subjects). Determination of the responsiveness statistic allows comparison among measures. This information also permits sample size calculations for future clinical trials.

A potential disadvantage of effect size and the Guyatt statistic as calculated is that the score change in the numerator may actually overestimate treatment effects. This may arise because some change is often noted even in stable patients. In order to adjust for this potential overestimation, the difference in score change observed in stable subjects should be subtracted from the numerator of either statistic. Subtracting this value from the numerator allows calculation of a revised effect size and revised responsiveness index which may be better estimates of the true treatment effect.

3.4: Sample Size Calculation:

In order to determine the sample size or number of subjects needed for future clinical trials, the two parameters used to calculate the responsiveness index can be used. These are the minimum clinically important difference and the variability in stable patients or ($\sqrt{}$(2MSE) (36). If the responsiveness of an instrument is known, one can choose the sample size required for an experiment where change in test score over time is the endpoint and in which pre and post treatment scores are available.

Generally the change in score that is the minimum clinically important difference is not known. This difference can be estimated by determining the change in score observed after an intervention of known efficacy. If a poor choice is made and the treatment does not work, responsiveness will be under estimated. There is no standardized method of estimating this value. Surveys of subjects particularly those who rate themselves as a "little improved" may provide better estimates of this value. Final estimation requires further information from future clinical trials and surveys of patient opinion regarding what and how much are important differences in their symptoms.

The second parameter required is the mean square error (MSE) or variability in stable patients. As mentioned in the previous section, the standard deviation equals the square root of twice the mean square error. The MSE can be calculated in the following manner. Simply squaring both sides of the equation results in the variance

equal to twice the MSE. Dividing each side of the equation by 2 gives the formula for MSE = $s^2/2$ where s is the standard deviation or s^2 the variance.

Knowing the responsiveness index (Guyatt statistic) allows determination of sample sizes required for further trials using the evaluated instrument. If the variability in stable patients is small relative to the subject score which constitutes a clinically important difference or treatment effect, a clinical trial could be conducted with small numbers. Guyatt has demonstrated such calculations and generated an illustrative table (see Table 3.1 below). The formulae used for sample size calculation for independent groups and for related groups are as follows: $2[(Z\alpha + Z\beta)\sigma/\Delta]^2$ and $[(Za + Z\beta)\sigma/\Delta]^2$ where α is the probability of erroneously concluding the treatment is effective, β is the probability of erroneously concluding the treatment is ineffective, Δ is the minimum clinically important difference, and σ the variability in stable patients or $\sqrt{}$(2xMSE) (36). Note that in Table 3.1, n is increased according to the method of Lachin when sample size is less than 30 (53). This explains the odd numbers generated in this table.

Table 3.1: Sample Size Calculations Using Guyatt's Responsiveness Index: (assumptions: $\alpha = 0.05$ (1-tailed); $\beta = 0.10$)

Guyatt Responsiveness Index ($\Delta/\sqrt{2}$xMSE)	Sample size required for independent groups	Sample size required for related groups
2	7	5
1	19	11
0.8	29	16
0.6	48	26
0.5	68	34
0.4	107	54
0.2	428	214

adapted from (36).

In the above table of sample size calculations, a one-tailed alpha was used. This assumption is reasonable in the instance where the investigator is interested only in detecting change in one direction such as when using a health instrument to only detect symptom improvement after treatment. The investigator would consider worsening of symptoms the same as no response to treatment. Conventionally most studies are designed to detect improvement or worsening of the outcome being measured. This calls for use of a two-tailed alpha which will increase the necessary sample size. Sample size calculations using a two-tailed alpha are illustrated in Table 3.2 below. The value of n is increased according to Lachin's method as in the previous example.

Table 3.2: Sample Size Calculations Using Guyatt's Responsiveness Index: (assumptions: $\alpha = 0.05$ (2-tailed); $\beta = 0.10$)

Guyatt Responsiveness Index ($\Delta/\sqrt{2}$xMSE)	Sample size required for independent groups	Sample size required for related groups
2	8	6
1	23	13
0.8	34	18
0.6	58	31
0.5	84	42
0.4	132	66
0.2	524	262

Chapter 4: NON-ULCER DYSPEPSIA:

4.1: Introduction:

The functional bowel disorders are a group of gastrointestinal conditions for which no structural abnormality is known to be the cause. These are important entities because of the frequency with which they occur in the general population. Nonulcer dyspepsia, one of the functional bowel disorders, is defined as chronic or recurrent (greater than three months duration) upper abdominal pain or nausea which may or may not be related to meals (54). The cause of nonulcer dyspepsia is unknown although a variety of abnormalities such as delayed gastric emptying, hypersensitivity to gastric distention, and *Helicobacter pylori* infestation are found in variable numbers of patients (54, 55).

The costs to the health care system related to nonulcer dyspepsia are staggering. The point prevalence of dyspepsia (upper abdominal pain or discomfort) has been estimated to be 25% with an annual incidence of 2-8% (56). Up to 5% of primary care visits are for dyspepsia of which 60% have no organic explanation. The majority of dyspeptic patients will receive a prescription. Patients with functional dyspepsia are two to three times more likely to be off work for health reasons (57).

Patients diagnosed with nonulcer dyspepsia have a variety of complaints. These may include abdominal pain, nausea, bloating, early satiety and retching. Because of the diversity of complaints

suffered by nonulcer dyspepsia patients, it is difficult to conceive of one unifying hypothesis to explain this condition. In fact, nonulcer dyspepsia may represent a number of different conditions. In an attempt to clarify and categorize nonulcer dyspepsia patients with a view to aiding therapy, a number of subgroups have been proposed. In 1988, an international working group proposed a classification of nonulcer dyspepsia subgroups based on symptoms (see Table 4.1, appendix) (58). These criteria have been more recently reviewed and updated (59).

Based upon this classification of nonulcer dyspepsia patients, therapy has been directed towards the specific symptom subgroup. Patients considered to fall in the ulcerlike subgroup have received therapy directed to acid suppression with agents such as the H_2 blockers or proton pump inhibitors. Patients suffering from motility like symptoms have been treated with prokinetic agents such as Cisapride or Domperidone. The available evidence suggests this is a reasonable approach (60-65).

4.2: Critical Appraisal of the Literature:

A large number of clinical studies have been performed and published in an attempt to determine potential treatments for patients suffering with nonulcer dyspepsia. To date, there is no definitive evidence for an efficacious treatment.

A number of criticisms of the current nonulcer dyspepsia literature have been put forward by Veldhuyzen van Zanten et al. in a recent systematic review (66). The following information is condensed from Veldhuyzen van Zanten's work.

Fifty-two randomized trials were identified of which 36 were placebo-controlled trials and 16 cross-over design studies. A variety of symptoms were measured in these trials. Forty-nine trials (94%) measured epigastric pain, 38 (73%) trials nausea, 29 (56%) trials heartburn, 25 (48%) trials bloating, and 28 (35%) trials belching. Categorical scales were the most common type of outcome measure used in 34 (65%) trials with use of symptom severity the most widespread. Thirteen (25%) trials assessed symptom frequency and only 5 (10%) estimated symptom duration. Four point categorical scales were the most popular. Only 5 studies used 5 or 7 point scales. Visual analog scales were used in 6 (12%) studies.

Patient based global assessments of overall symptom severity were used in 19 (37%) of the identified studies. In 14 of these 19 trials, the subjects were asked to make a global assessment of change in symptoms after the treatment had been received, without a baseline

assessment before the intervention. Eight trials (15%) used a physician-based global assessment as an outcome measure.

The major flaw in this body of literature was the lack of validation of outcome measures prior to their use in a clinical trial. Only 5 studies used outcome measures which had been previously validated in a pilot study (67-71). Of these 5 trials, 4 were multiple cross-over trials. To date, there has only been one placebo-controlled, not cross-over trial using previously validated outcome measures in nonulcer dyspepsia (67).

Veldhuyzen van Zanten and others have made a number of suggestions for future research in this area. Most studies do not clearly state whether enrolled patients were obtained from primary or tertiary practice. The patient setting should be clearly described to be aware of potential problems with patient selection and referral bias. A variety of definitions of nonulcer dyspepsia have been used, affecting reproducibility of results. It has been suggested by Veldhuyzen van Zanten that the Rome criteria for diagnosis of functional dyspepsia be adopted (59). Additionally, these experts have advised that patients diagnosed with the Irritable Bowel syndrome (IBS), as assessed by three or more of the Manning (72) or Rome criteria for IBS (73), be excluded from future studies of nonulcer dyspepsia.

There is consensus that a thorough workup is required. Nonulcer dyspepsia symptoms mimic other organic conditions. The diagnosis of NUD is one of exclusion. A minimum set of investigations should

include endoscopy and basic laboratory screening. Patients must have symptoms severe enough to seek medical attention and must still be symptomatic at the time of enrollment.

A set of exclusion criteria has also been recommended by Veldhuyzen van Zanten and his co-authors. These include the following: 1) history of, or evidence of esophagitis; 2) history of, or presence of gastric or duodenal ulcer; 3) endoscopic evidence of gastric erosions; 4) endoscopic duodenal erosions; 5) history of previous upper GI surgery; 6) daily use of a nonsteroidal anti-inflammatory drug (NSAID); 7) suspected or known alcoholism; and 8) presence of the irritable bowel syndrome.

Perhaps the most important recommendation is that all future trials should be randomized, placebo-controlled trials. The placebo response in NUD trials varies from 13 to 73%. Cross-over designs are not recommended for research in functional dyspepsia. In studies of cross-over design, each subject receives both treatments being compared in the clinical trial. Cross-over designs allow comparison of within patient differences rather than between patient differences of placebo-controlled trials. Consequently smaller sample sizes are required. Cross-over studies call for stable, usually chronic disease during both treatment periods and a similar baseline condition present at the start of each treatment period. If the patient baseline differs markedly at the start of each treatment, it is impossible to compare the two treatments. Finally, there should be no carry over effects after either treatment. This means that all disease manifestations revert to baseline and all the effects of previous

treatment disappear after cessation of therapy (74). Since the variability of symptoms in NUD can be substantial and because it is uncertain if patients will go back to baseline after a wash-out period post-treatment, Veldhuyzen van Zanten and co-authors have recommended that cross-over designs be avoided (66).

Another problem in this area of research is the tremendous variation in the use of outcome measures, including which symptoms are assessed and how symptom severity is measured. This same group of authors have recommended the use of 5 to 7 point categorical scales. These scales are more responsive than smaller scales. There has been no agreement how to measure symptoms, however an assessment of severity must be done at baseline before any intervention and symptoms must be of sufficient severity to allow documentation of any improvement. Patients with mild dyspepsia have little room to improve even with a truly effective drug. An overall subject-based global assessment of symptom severity should be included rather than using multiple symptoms to avoid potential problems with multiple comparisons. Physician-based assessments of overall symptom severity should not be used as the main outcome measure because there may be substantial inter-and intra-observer variation in physician recording of symptom severity. Physician assessments may be useful as secondary outcomes.

The final recommendation of this paper is that all subjective measures used to determine outcome must be validated prior to use in a clinical trial. This process involves the demonstration of reliability, responsiveness, and validity of the measure (66).

4.3: Available Outcome Measures:

The major weakness of this body of literature, as already stated, is the almost complete lack of use of validated outcome measures. By the very subjective nature of the functional bowel diseases, quantification of response to therapy is extremely difficult. The main focus of the nonulcer dyspepsia literature has been the establishment of the diagnosis of nonulcer dyspepsia. In recent years, investigators have taken some interest in the development and validation of outcome measures to evaluate symptom severity (66). To date, there are two fully published reports describing symptom-based outcome measures (75, 76) as well as a condition-specific measure which has not yet been fully validated (77).

Nyren et al developed a multidimensional symptom score for epigastric pain labeled the DIBS (duration-intensity-behaviour scale) (75). The DIBS scale was a seven point adjectival scale. This scale was developed by comparing epigastric pain response to antacid therapy for three weeks in 32 patients with functional epigastric pain. The pain index (the product of pain intensity and duration) in the DIBS was compared to a concurrently administered visual analog scale. A high degree of concordance among the scales was demonstrated and both were determined to be sensitive to change.

The DIBS has been used subsequently in a randomized controlled trial of 3 weeks of treatment with Cimetidine, antacid, or placebo in 159 nonulcer dyspepsia patients. The primary outcome was decrease

of epigastric pain. No significant difference was demonstrated between the three groups studied (67).

Nyren's paper has been criticized for the use of only one symptom (epigastric pain). Patients with nonulcer dyspepsia may suffer from a variety of other complaints such as nausea or bloating. It is not clear if epigastric pain alone is a sufficient indicator of functional dyspepsia. However, by focusing on a single symptom, the multiple comparisons seen in many nonulcer dyspepsia trials were avoided. It is unreasonable to assume that the symptoms of nonulcer dyspepsia are independent of each other. Using multiple symptoms as separate outcomes increases the possibility of concluding active treatment is superior to placebo simply by chance alone. This can be avoided by correcting for multiple comparisons by using statistical corrections such as the Bonferroni or Tukey corrections (78, 79) or instead by using a global measure of symptom severity as the primary outcome measure (66).

Veldhuyzen van Zanten published a symptom-based outcome measure for nonulcer dyspepsia patients following the guidelines established by Guyatt (26) for the development of disease-specific measures (76). This group looked at patients with *Helicobacter pylori* -associated gastritis and nonulcer dyspepsia (*Helicobacter pylori* negative patients). The objective of this work was to select gastrointestinal symptoms and establish that these symptoms recorded as 5-point adjectival scales would meet criteria for use as outcome measures in clinical trials. Symptoms were selected from the literature and a pilot group of 24 patients were used to reduce

the number of items for the questionnaire. The 8 symptoms with the highest cumulative scores (the product of severity and frequency) were selected for use. A second group of 55 patients was studied to test reproducibility, responsiveness, and validity of the preferred items. Patients with the irritable bowel syndrome and non-steroidal anti-inflammatory drug users were excluded, as was anyone with abnormal endoscopy. *Helicobacter pylori* status was determined. Symptom severity was measured at study entry (T1), at one week before treatment was given (T2), and at 4 weeks after treatment (T3). *Helicobacter pylori* positive patients received Pepto-Bismol and Ampicillin and Helicobacter pylori negative patients received antacids or H_2 blocker therapy. Reproducibility was tested by comparing repeated measurements before intervention (T1 and T2). Responsiveness or ability to detect change was assessed by comparing scores immediately before (T2) and after (T3) treatment. Validity was determined by comparing scores with changes in general health status measured by patient global assessments. The authors concluded that scoring gastrointestinal symptoms using 5-point adjectival scales satisfied the 3 criteria for use as outcome measures.

This study was methodologically very sound, however it is not without problems. Three of the 8 symptoms selected (heartburn, sour taste, and halitosis) are symptoms of acid reflux, not nonulcer dyspepsia. At the time this study was ongoing, patients with symptoms of acid reflux and normal endoscopy were considered to represent a reflux-like subgroup of nonulcer dyspepsia. Veldhuyzen

van Zanten, Talley, and others have since convincingly argued against the existence of this subgroup and have stated that patients with heartburn as the predominant symptom should not be included in studies of nonulcer dyspepsia (66). Many patients with pathologic acid reflux do not have endoscopic evidence of esophagitis (54, 59). A study has demonstrated that heartburn and acid regurgitation as the dominant symptoms are very specific (85% and 96%) for a diagnosis of gastroesophageal reflux disease (GERD) (80). In other words, if these symptoms are present the patient is likely to have acid reflux disease. Therefore it has been recommended to exclude patients whose predominant symptoms are heartburn or acid regurgitation since they have GERD, not NUD (66). Following this argument, many of the symptoms used in Veldhuyzen van Zanten's outcome measure should be excluded.

The most recent publication in this area is that of an ongoing development of a condition-specific questionnaire for dyspepsia and ulcer-related symptoms (77). Items have been selected by literature review with analysis by content experts and pretested with a small patient sample. Testing was performed by mailing the new questionnaire and a previously developed generic questionnaire (SF-36) (81) to identified referral patients from primary care. Initial items were rejected if too much correlation with other items was documented. In other words, little further information would be added with inclusion of these items in the questionnaire. A subsample received the revised second questionnaire along with the global assessment to test reproducibility. Validity was tested by

comparing the new questionnaire with the concurrently administered generic health survey. Responsiveness has not been tested to date with further questionnaire development ongoing.

This questionnaire included items regarding epigastric pain and associated features (location, severity, frequency) as well as symptoms such as vomiting, melena, hematemesis, and weight loss. Heartburn had been excluded. Many of these symptoms would not be seen in nonulcer dyspepsia patients. In fact, their very presence would be indicative of serious organic disease and consequently exclude the diagnosis of NUD. In addition, other symptoms commonly seen in nonulcer dyspepsia are not included, for example, many features of dysmotility including nausea and bloating. Consequently, this questionnaire is not appropriate for use in nonulcer dyspepsia trials.

In conclusion, no appropriate outcome measure exists for use in clinical trials of nonulcer dyspepsia. Therefore, a new symptom-based measure is required.

Chapter 5: OBJECTIVES:

5.1: Study Objective:

The main objective of this study was to develop a symptom-based outcome measure for future use in nonulcer dyspepsia trials. The outcome measure was designed using the recommendations of Veldhuyzen van Zanten and others for NUD trials. As part of this process, this measure would have to be shown to be reproducible, responsive, and valid.

Chapter 6: METHODS:

6.1: Questionnaire Development:

The questionnaire used in this study was developed before commencement of the study (see appendix). The principles suggested by Guyatt were used to guide questionnaire development (26). The questionnaire was intended to be short, simple, and easy to apply. It was directed specifically at symptoms known to be associated with nonulcer dyspepsia. The symptoms selected were abstracted from clinical guidelines in widespread use (58) and included abdominal pain, nausea, retching, vomiting, bloating, and early satiety.

Using a variation of a previously validated method by Nyren (75), each study participant was asked to select their most important symptom from the list provided. Nyren used only abdominal pain as the marker for response to therapy in nonulcer dyspepsia patients (75). His primary outcome measure was the product of abdominal pain intensity and duration (combined score). Since only using abdominal pain excludes a number of nonulcer dyspepsia patients, in particular the subgroup with motility symptoms such as bloating, nausea, or early satiety, the participants in this study were asked to select their most significant complaint or symptom. The selected complaint was used as the individual's indicator of response.

For this study, the frequency and severity of the selected symptom were recorded on 5-point adjectival scales. Scores were assigned as follows: none= 1 point, once per week/mild= 2 points, most

days/moderate= 3 points, daily/severe= 4 points, and more than once each day/very severe= 5 points. Five point scales were used since this outcome measure has previously been demonstrated to be reproducible and responsive when measuring individual gastrointestinal symptoms (76). The frequency/severity product was calculated and recorded as a estimate of the participant's present condition. Using the product score increased the range of possible results from 1 to 25. Since both frequency and severity were measured using the product, changes in either would be reflected in the final score. This would be expected to result in increased sensitivity to change of the patient's symptom. Perceived changes only in severity but not frequency would be considered a change for the patient but would not be detected by an instrument measuring frequency only. The product of frequency and severity of the subject-selected symptom was the primary outcome measure of this study.

A patient-based global assessment of overall status was recorded with each administration of the questionnaire using a 10 cm. visual analog scale. A VAS was chosen because of its simplicity, documented comparable responsiveness to categorical scales (34), and avoidance of potential 'halo' effect. The halo effect indicates the tendency of questionnaire recipients to fill in the same position on a scale when categorical scales are listed above and below each other on a page (1).

Other measures obtained included frequency of antacid administration and physician assessment of subject's symptom

severity. Antacid use, which was hypothesized to decrease if the subject improved, was recorded each time with a 5-point adjectival scale (not at all; once per week or less; most days but not every day; once per day; and more than once each day). The physician's impression of the severity of the patient's complaints was recorded with 5-point adjectival scales (none; minimal; mild; moderate; and severe). Five point scales were chosen to correlate with the previously mentioned symptom severity and frequency measures. Smaller scales were not used to avoid potential loss of information and sensitivity to change inherent in smaller scales.

Finally, in order to distinguish which patients had responded to their therapy, symmetrical 7-point adjectival scales were used. Seven point scales were selected because they should be more sensitive to change than 5-point scales. Increased sensitivity to change was felt to be important. In order to determine the responsiveness of the instrument, patients who had changed with therapy had to be selected from those who were unchanged. Physician-patient consensus was required to determine the changed group of subjects. Both the physician and patient indicated whether the participant had improved, had remained the same, or had worsened. The physician was unaware of the patient's response at the time of assessment.

6.2: Study Entry:

The patients in this study were referred from a general practice setting to a tertiary care gastroenterology clinic. Those people who complained of upper abdominal pain/discomfort for greater than six months were approached about entering this study. A routine history and physical examination was performed. Patients were excluded if they had undergone prior gastric surgery, might be pregnant, or had used proton pump inhibitors, antibiotics, PeptoBismol, or non-steroidal anti-inflammatory drugs (NSAID) in the previous month. Patients with symptoms of reflux disease such as heartburn, regurgitation, or water brash, or who required ongoing NSAID therapy were also excluded. Finally, patients diagnosed with the irritable bowel syndrome by the presence of three or more of the Manning Criteria (see Table 6.1) were not considered appropriate for this study. Subjects who gave informed consent were entered into the study.

An endoscopy was performed as per standard practice to rule out organic disease as the cause of the patients symptoms and to confirm the diagnosis of nonulcer dyspepsia. The presence of abnormal endoscopic findings excluded the subject from further study.

Once the diagnosis of nonulcer dyspepsia had been established, each participant was categorized into either the ulcerlike or dysmotilitylike subgroup based on the Rome criteria (58, 59). The presence of three or more criteria (Table 4.1) was required. Participants who had overlapping symptoms were classified in the

category for which the most criteria were present. The clinical nonulcer dyspepsia subgroup was used to guide the choice of intervention. Previous literature has documented the responsiveness of the ulcerlike subgroup to H_2-blockers (62-65) and the dysmotilitylike subgroup to motility drugs (60, 61). Members of the ulcerlike subgroup received acid suppressive therapy with H_2-blockers or proton pump inhibitors. The actual drug selection was left to the discretion of the investigator. The dysmotilitylike subgroup received a prokinetic agent, Cisapride or Domperidone, at the choice of the investigator. Treatment was given in standard doses for a period of one month. It should be noted that this was an open, uncontrolled study with no attempt to assess the benefit of treatment with one drug compared to another. The suitability of drug selection for the different subgroups was not assessed nor were the response rates of the subgroups compared.

6.3: Questionnaire Administration:

The first administration of the questionnaire (T1) was performed at the initial visit. Patients were instructed on how to fill in the questionnaire, with particular emphasis on how to record information using adjectival scales, and on the proper method of indicating overall status using a visual analog scale. The questionnaire was self-administered by the study participant without receiving further aid or intervention by the investigator. The investigator also recorded their own assessment of the severity of the subject's condition without prior knowledge of the participant's responses.

Subjects were given a second copy of the same questionnaire and asked to fill this in at home one week later (T2). The completed questionnaires were mailed in self-addressed, stamped envelopes which were provided. No intervention was given in the week between the first and second administration of the questionnaire. The data gathered at this point from untreated, presumably stable subjects, allowed assessment of the instrument's reliability.

At this point, the assigned therapy was begun for a one month period. After one month, each study participant returned to the GI clinic for assessment. A third copy of the questionnaire was self-administered (T3). A subject-based global assessment was again obtained on a visual analog scale and the subject also indicated whether they had any response to therapy using a 7-point adjectival scale. A physician assessment of the severity of the patient's

condition and of the subject's response to therapy were recorded by the investigator, again without prior knowledge of the participant's responses. The data obtained at this point permitted determination of the instrument's responsiveness to change in subjects who may have responded to treatment. (See appendix, third section).

6.4: Questionnaire Correlations:

To evaluate the validity of the new measure, guidelines for strength of correlation were modeled upon validation studies of 2 other health measures. In the validation of the Inflammatory Bowel Disease Questionnaire (IBDQ) (39), items were considered to be closely related when correlation > 0.5. For example, the authors of this instrument felt that the patient global rating of change in disease activity would be closely related to the change in bowel symptoms dimension of the IBDQ, Measures were considered moderately related when correlation > 0.4. In this instance the authors predicted that the relative's global rating of disease activity would relate moderately well to change in the bowel symptom dimension. Some relationship was considered if correlation > 0.3.

The Asthma Quality of Life Questionnaire was a similar instrument developed for use in patients with asthma (82). In this study, the authors used the following indices to determine strength of relationship: strongly correlated, $r > 0.5$; moderately correlated, $r = 0.35$ to 0.5; and poorly correlated, $r = .20$ to .35.

In the present study, a variety of constructs were predicted to be related to the new measure. Based upon information from the two previously mentioned validity trials, it was expected that patient global assessment should be closely related to the frequency-severity product of the subject-selected symptom. Physician global assessment should show moderate relationship with the frequency-severity product. There was no previous data for comparison

regarding the correlation of antacid use with response to treatment in nonulcer dyspepsia patients, however it was felt that there should be some relationship. Antacid use should decline in subjects who have responded to the treatment received.

6.5: Statistical Analysis:

Results were analyzed using a variety of statistical methods. Ordinal data generated by categorical scales were analyzed using nonparametric methods such as Spearman Rank Correlation (46). P-values less than 0.05 were considered to be statistically significant. The Intraclass Correlation Coefficient was also calculated to assess the instrument's reliability. Effect size and the Guyatt Responsiveness statistic were calculated to assess the instrument's responsiveness (50). See appendix for details of the methods of calculation of these statistics. Statistical calculations were performed using Statview v.4.5 for the Macintosh (83) on a Power Macintosh 6100 computer.

Chapter 7: RESULTS:

7.1: Characteristics of Study Participants:

Fifty-five people were approached for this study between October 1994 and November 1996. Two people refused to participate when asked to enter the study. Another 9 who had given consent were excluded after endoscopy because of abnormal endoscopic findings (4 with duodenal ulcers, 2 with esophagitis, 1 with erosive gastritis, and 2 with erosive duodenitis). Forty-four participants were enrolled in this study and all 44 completed the first questionnaire. Thirty-eight (86%) completed the second questionnaire T2 and 35 (80%) completed the third and final questionnaire T3. Eighteen of 21 in the ulcerlike group (86%) and 17 of 23 (74%) of the dysmotilitylike group completed the study. Characteristics of the study participants are listed in Table 3 below. The participants in the two subgroups were similiar with the exception of smoking which was more prevalent in the dysmotilitylike group.

Table 7.1: Patient characteristics:

Attribute	Ulcerlike subgroup	Dysmotility -like subgroup	Total	p-value
Number of participants	21	23	44	ns
Mean Age	35.5	40.5	44	ns
Female Gender	14	17	31	ns
Smoker	1	7	8	.048
Alcohol User	3	4	7	ns
Other Medications	7	5	12	ns
Caffeine User	16	20	36	ns

7.2: Reproducibility Statistics:

To determine reproducibility, assessment of stable patients on at least two separate occasions was required. T1 and T2 assessments were performed one week apart in patients who had received no intervention and thus should have been stable clinically. The frequency-severity products obtained with the first (T1) and second (T2) administrations of the questionnaires are recorded in Table A1 (see appendix). The range of possible scores for Frequency and Severity ratings are 1 to 5. The frequency-severity product scores ranged from 1 to 25. A comparison of the scores obtained at T1 and T2 was performed using Spearman Rank Correlation. The rho values are listed below in Table 7.2.

Table 7.2: Comparison of symptom scores obtained 1 week apart before therapy received.

Category	T1 mean score	T2 mean score	Spearman coefficient (95%CI)	p-value
Main Symptom Frequency	3.6	3.8	.73 (.62, .84)	<0001
Main Symptom Severity	3.7	3.5	.86 (.78, .94)	<0001
Frequency/ Severity Product	13.5	13.4	.85 (.76, .94)	<0001

Because of the concern that the Spearman coefficient concentrates on between subject variability rather than measuring within subject variation the Intraclass Correlation Coefficient was also calculated (50). See appendix for detailed description of calculation of this statistic. The calculated ICC's for symptom scores are listed in Table 7.3 below.

Table 7.3: Intraclass Correlation Coefficients (ICC).

Symptom Category	Standard Deviation of T1 (A)	Standard Deviation of T2 (B)	Standard Deviation of T1-T2 (C)	Intraclass Correlation Coefficient (95%CI)
Frequency	1.15	1.00	0.79	0.73 (.62, .84)
Severity	0.87	0.76	0.44	0.84 (.75, .93)
Product	5.90	5.23	3.29	0.83 (.74, .92)

7.3: Responsiveness Statistics:

As mentioned earlier, a variety of statistical methods have been used to determine the responsiveness or ability of an instrument to detect change. These methods include t-tests to compare sample means, indicators of effect size, and the responsiveness index, a modified effect size statistic proposed by Guyatt. T-tests were not used in this study since there was no other subjective outcome measure available for comparison.

In order to be able to assess the responsiveness of an instrument, evaluation needs to be performed on patients who have improved upon receiving an intervention of known efficacy. The data used in the following analyses was obtained from participants who were considered to have responded to the therapy they had received. The 27 participants who had improved as indicated by both patient and physician global assessments were selected. Fourteen of 18 (78%) of the ulcerlike subgroup were judged to have improved with treatment as did 13 of the 17 (76%) in the dysmotilitylike subgroup. Of the eight subjects judged not to have improved, only one indicated improvement by patient assessment but not physician assessment. The remaining seven were unchanged by both assessments. The resulting data (see Table A 2, appendix) was analyzed to determine the effect size and Guyatt responsiveness statistic which are listed in Table 7.4 below. It should be noted that three subjects #9, #22, and #29 did not complete the mail-in questionnaire (T2) but did complete the final questionnaire (T3) after receiving treatment.

Consequently these subjects were not used in the calculation of reliability (Table A1) but were used to calculate instrument responsiveness (Table A2).

Table 7.4: Responsiveness Statistics.

Variable	Effect Size	Guyatt Statistic
Frequency	1.5	1.2
Severity	1.4	1.1
Product	1.3	2.1
Revised Product	1.2	2.1

From the data one can see that each variable assessed was quite responsive to change. The frequency-severity product seemed to be the most sensitive to change, more than either symptom frequency or severity.

As previously mentioned, the effect size and the Guyatt statistic as calculated may actually overestimate treatment effects. To adjust for this the difference in score change observed in stable subjects was subtracted from the numerator of each statistic. The score change in stable patients in this study (before the intervention was received) was calculated to be 0.1 (see Table A1, appendix). Subtracting this value from the numerator allowed computation of a revised effect size of 1.2 and a revised responsiveness index of 2.1 (see Table 7.4 above).

7.4: Validity:

In order to validate a new symptom-based outcome measure, the newly developed instrument needs to be assessed, preferably by criterion validity. Criterion validity involves comparison of the new measure against a reference measure that evaluates the same or similar features (28). This might, for example, consist of comparing the new instrument to a generic measure. There was no accepted generic measure of this type in nonulcer dyspepsia available for comparison. Comparison to other disease-specific instruments and symptom-based outcome measures was not performed because of the limitations of the existing instruments as previously discussed. These limitations included exclusion of certain nonulcer dyspepsia symptoms and inappropriate inclusion of gastroesophageal reflux disease patients.

Since demonstration of criterion validity was not an option, construct validation was used. Construct validation involves comparison of changes in the new instrument with changes in other measures including subjective assessments by a physician, relative, or the patient. If the measure is valid, improvements in the subject's status as indicated by the frequency-severity product should correlate with improvements as indicated by the constructs. To demonstrate construct validation of this instrument, the subject-selected symptom was compared with patient global assessment, physician global assessment, and assessment of the subject's antacid use.

The difference in pre- and post-treatment patient global assessments ranked by visual analog scale was assessed and compared with change in instrument scores of the frequency-severity product. The correlation coefficient r= .596 (p= .0005) indicated that patient global assessment by visual analog scale and the frequency-severity product were closely related.

The second construct assessed was a patient assessment of change in status after treatment recorded on a 7-point adjectival scale. It should be noted that higher frequency-severity product score differences indicated improvement whereas lower scores on the adjectival scales indicated improvement. The Spearman rank correlation with change in instrument scores pre- and post-treatment was r= ⁻.584 (p= .0007). The negative correlation coefficient in this instance indicates that high values of one variable tested (product differences) correlated with low values of the other variable (adjectival scale) (84). The assignment of values to the adjectival scales could be arbitrarily changed to give a positive but equivalent correlation coefficient. In this case, high product values would correlate with high values on the adjectival scale. The r value obtained by this comparison indicated the frequency-severity product and patient assessment of response to therapy by adjectival scale were closely related.

The third construct used for comparison was a physician global assessment of the subject's status. The physician assessment was compared to the change in instrument scores before and after treatment. The correlation coefficient r= ⁻.437 (p= .0088) indicated

some correlation between the instrument and physician assessment of patient status. In this instance, high instrument scores (product differences) correlated with low physician assessment scores.

The next construct assessed was a physician assessment of change in status after treatment, recorded on a 7-point adjectival scale. Some correlation with the difference in frequency-severity product was detected with $r = -.329$ ($p = .055$).

The final construct utilized was the amount of antacid consumed before and after the intervention. Analysis was performed only on the subjects who used antacids during the study time period. The correlation coefficient obtained by comparing change in antacid use with change in instrument scores $r = -.143$ ($p = .327$). This result indicated that there was no correlation of instrument scores with changes in antacid use in the subjects using antacids.

Chapter 8: DISCUSSION:

8.1: Introduction:

The purpose of this study was to develop and test a new symptom-based outcome measure for future use in therapeutic trials of nonulcer dyspepsia patients. In order to the confirm the adequacy of this measure, reproducibility, reliability, and validity of the instrument had to be evaluated. A new subjective outcome measure was thought necessary because of the overall lack of validated subjective outcome measures in the nonulcer dyspepsia literature and the limitations of the pre-existing measures such as inclusion of subjects with gastroesophageal reflux and other conditions incompatible with a diagnosis of NUD.

8.2: Questionnaire Design and Administration:

This questionnaire was designed following published guidelines for developing symptom-based outcome measures in nonulcer dyspepsia. Symptom severity and frequency were measured by 5-point scales which have been previously shown to be a valid method of measuring gastrointestinal symptoms. The physician assessment of subject status and antacid use were measured with 5-point scales. Response to therapy as indicated by subject and physician were recorded on 7-point scales. The rationale for this design has been described. These items could have been designed as 7-point scales. This alteration might have improved correlation particularly of the physician assessment, although the moderate correlation noted in this study (r= ⁻.437) was in keeping with physician-based assessments in other published studies. Any changes to the design of the questionnaire would require repeat testing in a pilot study prior to use in a clinical trial.

A visual analog scale was used to measure patient global assessment. Although a different instrument was used, the end result was similar to those obtained with categorical scales. The correlation of the VAS with the frequency-severity product (r= .596) and the correlation of response to treatment measured with a 7-point scale to the product (r=. ⁻584) were almost identical.

The questionnaire was administered at entry (T1), one week later before any treatment had been received (T2), and again after treatment for one month (T3). The reasons for the timing of

questionnaire administration have been described above. Subjects were allowed to perform T2 administration at home without coming in to the GI clinic. The study was purposely planned in this fashion to reduce the number of clinic visits. Many of the potential participants came from long distances. Reducing travel time was felt necessary to aid recruitment of subjects into the study.

Administration of the questionnaire at home could potentially lead to problems. For example, there was no way for the investigators to actually determine when the questionnaire was filled out. Secondly, the investigators could not definitively state that treatment had not been started before the second questionnaire was completed. In an attempt to avoid these problems, subjects were given an assigned date to complete the questionnaire and prescriptions for study drugs were pre-dated to start at the correct time. Data analysis suggests that significant effects did not arise from this design. The difference in the means of frequency-product scores at T1 (T1 mean score minus T3 mean score = 7.1) and T2 (T2 mean score minus T3 mean score = 7.3) were almost identical.

8.3: The Outcome Measure:

An adaptation of the previously validated technique by Nyren in nonulcer dyspepsia patients was used in this study. In this case, the main outcome measure was the frequency-severity product of the subject-selected symptom. The reasons for selection of this measure have been outlined.

One potential criticism of this measure is the use of 5-point scales for measuring symptoms. Seven point scales would presumably be more responsive to change, however this did not appear to be a problem since this instrument is quite responsive in its present format (see below). The use of the subject's most significant symptom poses another potential problem for this measure if subjects were to change the selected symptom part way through the assessment period. This did not occur during this study, as no subject changed their selected symptom at T2 compared to T1. This potential concern could be avoided in future by redesigning the questionnaire so that the frequency-severity product of all symptoms were measured. This would increase the complexity of the questionnaire and raise the issue of multiple statistical tests. The study participants were not asked to select their most important symptom at T3 since this might well have changed with treatment. In addition, to adequately assess responsiveness, the measured symptoms cannot be so mild that response to treatment cannot be measured. Using the subject's most significant symptom avoids this potential pitfall.

8.4: The Study Population:

Forty-four subjects diagnosed with nonulcer dyspepsia who met the inclusion criteria were enrolled in this study. Thirty-five (80%) completed the study as per the planned protocol. Follow-up of subjects who failed to complete the questionnaire was attempted by phone or mail. Dropouts were equal among the two clinical subgroups, suggesting that the reasons for leaving the study were not due to adverse events of one drug class or failure of one specific treatment (i.e. the prokinetics did not work in the dysmotilitylike subjects).

After diagnosis and enrollment each subject was classified into a clinical subgroup based upon predetermined criteria (see appendix, Table 4.1). Fifteen of the 44 participants had 1 or 2 symptoms compatible with the other subgroup into which they were not classified. Patients with overlapping symptoms were still placed into the appropriate clinical subgroup. Since the intent was to mimic routine clinical practice where NUD patients are treated with acid suppressive or prokinetic agents based upon the predominant symptom pattern, no subject was excluded. None of the 15 participants with overlapping symptoms had three or more symptoms from the other subgroup requiring categorization into both subgroups.

The characteristics of the study subjects are listed in Table 7.1. The subjects in each subgroup were similar except that the prevalence of smokers was higher in the dysmotilitylike subgroup. The reason for

this difference was not obvious but may be due to the small number of smokers in this study. One must also be cautious in interpretation of p-values since multiple comparisons increase the possibility of finding a statistically significant difference in one of the comparisons simply by chance alone. Correcting for multiple comparisons using Bonferroni's theorem (p-value for each comparison should be multiplied by the total number of comparisons) would result in a lower significant p-value (84), thus suggesting the observed difference of smokers between the subgroups is not statistically significant. A relationship between smoking and nonulcer dyspepsia has not been reported previously. In fact, two recent publications suggested that there was no association between smoking and specific symptoms or subgroups of nonulcer dyspepsia (85, 86).

8.5: Instrument Reliability:

Reliability of the instrument was assessed by comparing symptom frequency, severity, and product scores by Spearman rank correlation. Correlation coefficients greater than 0.7 were generated (Table 7.2). The correlation coefficient obtained using the frequency/severity product was 0.85. In other words, in stable or untreated nonulcer dyspepsia patients, similar symptom scores were obtained on repeated measurements one week apart. The product of symptom severity and frequency seemed to be a more reliable outcome measure than symptom frequency but equivelent to symptom severity.

As recommended, the Intraclass Correlation Coefficient was calculated to assess reliability. Calculating the ICCs confirmed the reliability of this instrument (see Table 7.3). Strong correlation (ICC > 0.7) was demonstrated. As noted with the Spearman rank correlation coefficient, the frequency/severity product was a more reliable means of assessing the participant's outcome than frequency and equivalent to severity. The ICC of the frequency/severity product was 0.83, equivalent to the Spearman correlation coefficient obtained.

The ICC relates between-subject variance to the total variance. ICC values range from 0 to 1. When ICC values approach 0, systematic or random differences between the baseline and follow-up scores are present. The ICC approaches 1 when the variability between subjects increases. If the ICC is high, as was seen with this instrument (0.81),

then not much of the variability is due to variability in measurement on different occasions. In other words, reproducibility is high (50).

8.6: Instrument Responsiveness:

The effect size was used to assess the responsiveness of the instrument. The effect size relates changes in mean score to the standard deviation of the baseline scores. Participants who improved (see Table A2, appendix) as indicated by an external criterion (physician-patient consensus) were used to determine the effect size. The results obtained (see appendix for method of calculation) disclosed an effect size of the frequency-severity product equal to 1.3.

The effect size transforms the score change into a unit of measurement. If other measures become available in future, or if this measure underwent further modification, the effect sizes of the new measures would allow direct comparison among measures to determine the most responsive instrument (50). The instrument with the largest effect size would be deemed the most responsive.

The responsiveness index calculated for the frequency/severity product was 2.1. This was greater than the values calculated for frequency or severity alone (table 7.4), suggesting the product was the more responsive measure. The responsiveness index of 2.1 indicated that variability in stable patients was quite small in relation to the change in subject score. This value suggested that the measure tested in this study (frequency/severity product of the subject-selected main symptom) was highly responsive to change in nonulcer dyspepsia patients.

The difference in test scores obtained before and after the therapeutic intervention was actually the treatment effect of the subjects in this study. To reliably estimate sample size requirements for future trials with the instrument in question, the minimally important clinical difference is required. This difference is said to be the minimum change at which the patient group in question would feel any benefit from the therapy they had received. This value is actually not known for the patients reported in this study or for nonulcer dyspepsia patients in general. The minimally important clinical difference is certainly less than the estimated treatment effect of this study. In an attempt to better estimate the minimally important clinical difference, the product change was determined in the nine subjects who rated themselves "a little better" (see Table A3, appendix). The difference in mean scores in this small subset of subjects was 2.7 compared to the overall study score difference of 7.3. The responsiveness index in this instance equaled 0.77.

Using the responsiveness index of this instrument 2.1, sample size calculations previously done by Guyatt (see table 3.1) suggest that very small samples of 5 to 10 patients per group would be required in future trials. If the responsiveness index of 0.77 generated from the subjects who changed a little is a truer estimate, the required sample sizes would be still be in the range of 30 to 40 subjects in each arm of the study. From this, one can conclude that the instrument used in this study was responsive to change in NUD subjects and that this instrument would be a reasonable outcome measure for a randomized clinical trial.

8.7: Instrument Validity:

Upon demonstration of the reliability and responsiveness of symptom-based outcome measure, the validity of the instrument should be confirmed. The use of criterion validity was not an option in nonulcer dyspepsia patients. Other types of validity had to be used to assess the validity of the instrument in this study.

Face validity was apparent. This instrument was simple and easy to use. The instrument only required a few minutes of the subject's time. No problems in filling out the questionnaire were reported to the investigators by study participants.

Content validity, or the extent to which the domain of interest is sampled by the instrument, was apparent. The appropriate subjects were entered into the study by strict adherence to recommended definitions of NUD, exclusion of organic disease in all potential subjects by endoscopy , and usage of all recommended exclusion criteria for NUD trials including non enrollment of IBS and GERD patients. Furthermore, the subject was permitted to select their most important symptom, without influence from the investigator. Although the symptom was selected from a list of typical nonulcer dyspepsia symptoms, the participant was not restricted to this list.

Construct validity was used to validate the instrument in this study. Subject-based global assessments and assessments of improvement after therapy demonstrated good correlation with the frequency-severity product ($|r| = .596$ and .584). It should be noted that the

results were similiar regardless of the type of measurement used(visual analog scale versus adjectival scale). These correlations were similar to patient-based assessments noted in the previously mentioned validation studies (39, 82).

Physician-based assessments including global assessments and assessment of improvement after therapy showed less correlation to the instrument. This was also in keeping with the experience in the previous validation studies. The physician global assessment was moderately correlated with the instrument (|r|=.437). Only some correlation was noted between the instrument and the physician assessment of change after treatment (|r|=.329).

The final construct assessed was change in antacid use by the participant. Comparing antacid use before and after the intervention with change in the frequency/severity product revealed little or no correlation (|r|= .143). The lack of correlation with change in antacid use was probably explained by a relative ineffectiveness of antacids for this condition. Antacid use by the nonulcer dyspepsia patients in this study was noted to be rare or occasional throughout the study. Only eighteen (51%) of the participants ingested antacids at all during the study period with most of those using antacids less than once per week. There were no obvious trends of antacid use based on clinical subtype with eight ulcerlike and ten dysmotilitylike subjects consuming antacids during the study. Finally, measuring antacid use may simply be a poor construct for comparison for other reasons that are not clear.

In summary, this questionnaire appears to be a valid measure of the response of nonulcer dyspepsia subjects to therapy. Further study with this measure will be required to better assess its true validity. Use of the instrument in randomized clinical trials and use in different patient groups should confirm the validity of this instrument.

Chapter 9: CONCLUSION:

9.1: Conclusion:

This study documented the development of a new symptom-based outcome which can be used for future trials of patients with nonulcer dyspepsia. Using techniques described by Guyatt, Deyo and others this outcome measure met the criteria for assessment of a new measure. In other words, this instrument which used a patient-selected symptom outcome was demonstrated to be reliable in stable patients, responsive to change in patients responding to therapy, and valid, fulfilling the required criteria for further use in clinical studies.

This instrument is suitable for application as a subjective symptom-based outcome measure in future therapeutic trials of nonulcer dyspepsia.

Bibliography:

1. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. Oxford, England: Oxford University,1989

2. Spitzer WO, Dobson AJ, Hall J, et al. Measuring the quality of life of cancer patients. J Chron Dis 1981;34:585-97.

3. Helewa A, Goldsmith CH, Smythe HA, et al. Independent measurement of functional capacity in rheumatoid arthritis. J Rheumatol 1982;5:794-7.

4. Goldman L, Hashimoto D, Cook EF. Comparative reproducibility and validity of systems for assessing cardiovascular functional class: advantages of a new activity scale. Circulation 1981;64:1227-34.

5. Mahler DA, Weinberg DH, Wells CK, et al. Measurement of dyspnea: description of two new indices, interobserver agreement and physiological correlations. Am Rev Respir Dis 1982;24 (suppl 1):138.

6. Larson JL. The measurement of health: concepts and indicators. New York: Greenwood Press,1991

7. Schipper H, Clinch J, Powell V. Definitions and conceptual issues. In: Spilker B, ed. Quality of life assessments in clinical trials. New York: Raven Press Ltd., 1990:14-22.

8. Testa MA, Simonson DC. Assessment of Quality-of-life Outcomes. NEJM 1996;334(13):835-40.

9. Guyatt GH, Feeney DH, Patrick DL. Measuring health-related quality of life. Ann Int Med 1993;(118):622-29.

10. Berger M, Bobbitt RA, Carter WB, et al. The Sickness Impact Profile: development and final revisions of a health status measure. Med Care 1981;19:787-805.

11. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med Care 1989;27: S217-32.

12. Ott CR, Sivarajan ES, Newton KM, et al. A controlled randomized study of early cardiac rehabilitation: the SIP as an assessment tool. Heart Lung 1983;12:162-70.

13. Liang MH, Larson MG, Cullen KE, et al. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-47.

14. Deyo RA, Diehl AK, Rosenthal M. How many days of bed rest for acute low back pain? A randomized clinical trial. NEJM 1986;315: 1064-70.

15. Guyatt GH, Veldhuysen Van Zanten SJO, Feeny DH, et al. Measuring quality of life in clinical trials: a taxonomy and review. CMAJ 1989;140:1441-48.

16. MacKenzie CR, Charlson ME, DiGioia D, et al. Can the SIP measure change? An example of scale assessment. J Chron Dis 1986;39:429-38.

17. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnositic test performance. J Chron Dis 1986;39:897-906.

18. Kaplan RM, Bush JW. Health-related quality of life measurement for evaluation research and policy analysis. Health Psychol 1982;1:61-80.

19. Spitzer WO. State of science 1986: quality of life and functional status as target variables for research. J Chron Dis 1987;40:465-71.

20. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, et al. The measurement of patients' values in medicine. Med Decis Making 1982;2:449-462.

21. Guyatt GH, Jaeschke R. Measurements in Clinical Trials: Choosing the Appropriate Approach. In: Spilker B, ed. Quality of Life Assessments in Clinical Trials. New York: Raven Press Ltd., 1990: 37-46.

22. Olsson G, Lubsen J, van Es G, et al. Quality of life after myocardial infarction: effect of long term metoprolol on mortality and morbidity. Br Med J 1986;292:1491-93.

23. Fries JF, Spitz P, G KR, et al. Measurement of patient outcomes in arthritis. Arthritis Rheum 1980;23:137-45.

24. Schipper H, Clinch J, McMurray A, et al. Measuring the quality of life of cancer patients: the Functional Living Index-Cancer: development and validation. J Clin Oncol 1984;2:472-83.

25. Guyatt GH, Berman LB, Townsend M, et al. A measure of quality of life for clinical trials in chronic lung disease. Thorax 1987;42: 773-778.

26. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. CMAJ 1986;134(April 15): 889-95.

27. Wiklund I, Karlberg J. Evaluation of quality of life in clinical trials: Selecting quality-of-life measures. Cont Clin Trials 1991;(12):204S-216S.

28. Irvine JE, Feagan B, Rochon J, et al. Quality of life: a valid and reliable measure of therapeutic efficacy in the treatment of inflammatory bowel disease. Gastroenterol 1994;(106):287-96.

29. Huskisson EC. Measurement of pain. Lancet 1974;(2):1127-31.

30. Aitken RCB. A growing edge of measurement of feelings. Proceed Royal College Med 1969;62:989-92.

31. Scott PJ, Huskisson EC. Accuracy of subjective measurements made with and without previous scores. An important source of error in serial measurement of subjective states. An Rheum Dis 1978;38:558-9.

32. Likert RA. A technique for the development of attitude scales. Educ Psychol Measurement 1952;12:313-5.

33. Talley NJ, Nyren O, Drossman DA, et al. Toward optimal design of controlled treatment trials, with special reference to the irritable bowel syndrome. Gastroenterol Int 1993;6:189-211.

34. Guyatt GH, Townsend M, Berman LB, et al. A comparison of Likert and visual analog scales for measuring change in function. J Chron Dis 1987;40(12):1129-33.

35. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley,1972

36. Guyatt GH, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171-78.

37. Kirshner B, Guyatt GH. A methodologic framework for assessing health indices. J Chron Dis 1985;(38):27-36.

38. Jenkinson C. Evaluating the efficacy of medical treatment: possibilities and limitations. Soc Sci Med 1995;41(10):1395-1401.

39. Guyatt G, Mitchell A, Irvine EJ, et al. A new measure of health status for clinical trials in inflammatory bowel disease. Gastroenterol 1989;96:804-10.

40. MacKenzie CR, Charlson ME. Standards for the use of ordinal scales in clinical trials. BMJ 1986;292:40-3.

41. Van Hees JE, Van Elteren PH, Van Lier HJJ, et al. An index of inflammatory activity in patients with Crohns disease. Gut 1980;21:279-86.

42. Powell-Tuck J, Bown RL, Lennard-Jones JE. A comparison of oral prednisone given as single or multiple daily doses for active proctocolitis. Scand J Gastroenterol 1978;13:833-7.

43. Ware JE, Brook RH, Davies-Avery A, et al. Conceptualisation and measurement of health for adults in the health insurance study. In: Model of health and methodology. Santa Monica, Calif: Rand Corp., 1980: vol 1.

44. Carmines EG, Zeller RA. Reliability and validity assessment. Beverley Hills, Calif: Sage Publications,1979

45. Pearson K, Lee A. On the laws of inheritance in man. I. Inheritance of physical characters. Biometrika 1903;2:357-462.

46. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15:72-101.

47. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-34.

48. Kramer MS, Feinstein AR. Clinical biostatistics LII: the biostatistics of concordance. Clin Pharmacol Ther 1981;26:111-23.

49. Kazis LE, Anderson JJ, Meenan RF. Effect size for interpreting changes in health status. Med Care 1989;27(3,suppl):S178-89.

50. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. Control Clin Trials 1991;12:142S-158S.

51. Boucet C, Guillemin F, Briancon S. Nonspecific effects in longitudinal studies: impact on quality of life measures. J Clin Epidemiol 1996;49(1):15-20.

52. Fletcher RH, Fletcher SW, wagner EH. Clinical epidemiology: the essentials. Baltimore: Williams and Wilkins, 1984:127-152.

53. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Controlled Clin Trials 1981;2: 93-113.

54. Talley NJ, Phillips SF. Non-ulcer dyspepsia: potential causes and pathophysiology. Ann Int Med 1988;(108):865-79.

55. Nyren O. Functional dyspepsia: Is gastric acid and/or Helicobacter pylori infection involved in the aetiology. Eur J Gastroenterol Hepatol 1992;(4):608-14.

56. Talley NJ, Weaver AL, Zinmeister AR, et al. Onset and disappearance of gastrointestinal symptoms and functional gastrointestinal disorders. Am J Epidemiol 1992;(136):165-77.

57. Nyren O, Lindberg G, Lindstrom E, et al. Economic costs of functional dyspepsia. In: Langhorne PA, ed. Pharmacoeconomics. Adis International Ltd., 1992:312-24.

58. Working Party Report. Management of dyspepsia. Lancet 1988;(i):576-9.

59. Talley NJ, Collin-Jones D, Koch KL, et al. Functional dyspepsia: A classification with guidelines for diagnosis and management. Gastroenterol Int 1991;(4):145-60.

60. Inoue M, Sekiguchi T, Harasawa S, et al. Dyspepsia and dyspepsia subgroups in Japan: symptom profiles and experience with Cisapride. Scand J Gastroenterol 1993;28(Suppl 195):36-39.

61. Jian R, Ruskone A, Chaussade S, et al. Symptomatic, radionuclide and therapeutic assessment of chronic idiopathic dyspepsia. Dig Dis Sci 1989;34(5):657-64.

62. Lance P, Wastell C, Schiller KFR. A controlled trial of cimetidine for the treatment of nonulcer dyspepsia. J Clin Gastroenterol 1986;8(4):414-18.

63. Nesland AA, Berstad A. Effect of cimetidine in patients with non-ulcer dyspepsia and erosive prepyloric changes. Scand J Gastroenterol 1985;20:629-35.

64. Saunders JHB, Oliver RJ, Higson DL. Dyspepsia: incidence of non-ulcer disease in a controlled trial of ranitidine in general practice. Br Med J 1986;292:665-68.

65. Talley NJ, McNeil D, Hayden A, et al. Randomized, double-blind, placbo-controlled crossover trial of Cimetidine and Pirenzepine in nonulcer dyspepsia. Gastroenterol 1986;91:149-56.

66. Veldhuyzen van Zanten SJO, Cleary C, Talley NJ, et al. Drug treatment of functional dyspepsia: A systematic analysis of trial methodolgy with recommendations for design of future trials. Am J Gastroenterol 1996;91(4):660-73.

67. Nyren O, Adami H, Bates S, et al. Absence of therapeutic benefit from antacids or Cimetidine in non-ulcer dyspepsia. NEJM 1986;314(6):339-43.

68. Johannessen T, Fjosne U, Kleveland P, et al. Cimetidine responders in non-ulcer dyspepsia. Scand J Gastroenterol 1988;23:327-36.

69. Johannessen T, Kristensen P, Petersen H, et al. The symptomatic effect of 1-day treatment periods with cimetidine in dyspepsia. Combined results from randomized, controlled, single-subject trials. Scand J Gastroenterol 1991;26:974-80.

70. Farup P, Larsen S, Ulshagen K, et al. Ranitidine for non-ulcer dyspepsia. A clinical study of the symptomatic effect of ranitidine and a classification and characterisation of the responders to treatment. Scand J Gastroenterol 1991;26:1209-16.

71. Kleveland PM, Johannessen T, Kristensen P, et al. Effect of pancreatic enzymes in non-ulcer dyspepsia. A pilot study. Scand J Gastroenterol 1990;25:298-301.

72. Manning AP, Thompson WG, Heaton KW, et al. Towards a positive diagnosis of the irritable bowel. Br Med J 1978;(2):653-4.

73. Drossman DA, Thompson WG, Talley NJ, et al. Identification of subgroups of functional gastrointestinal disorders. Gastroenterol Int 1990;(3):159-72.

74. Spilker B. Guide to clinical trials.New York: Lippencott-Raven Pub.,1996

75. Nyren O, Adami HO, Bates S, et al. Self-rating of pain in nonulcer dyspepsia. J Clin Gastroenterol 1987;9(4):408-14.

76. Veldhuyzen van Zanten SJO, Tytgat KMAJ, Pollak PT, et al. Can severity of symptoms be used as an outcome measure in trials of non-ulcer dyspepsia and Helicobacter pylori associated gastritis. J Clin Epidemiol 1993;46(3):273-9.

77. Garratt AM, Ruta DA, Russell I, et al. Developing a condition-specific measure of health for patients with dyspepsia and ulcer-related symptoms. J Clin Epidemiol 1996;49(5):565-71.

78. Godfrey K. Comparing the means of several groups. NEJM 1985;313:1450-56.

79. Tukey JW. Some thoughts on clinical trials, espcially problems of multiplicity. Science 1977;198:679-84.

80. Klauser G, Schindlbeck NE, Muller-Lissner SA. Symptoms in gastroesophageal disease. Lancet 1990;335:205-8.

81. Ware JE, Sherbourne CD. The SF36 health status survey. 1. Conceptual framework and item selection. Med Care 1992;30:473-83.

82. Juniper EF, Guyatt GH, Ferrie PJ, et al. Measuring quality of life in asthma. Am Rev Respir Dis 1993;147:832-8.

83. Abascus Concepts Statview. In: 4.5 ed. Berkeley CA: Abascus Concepts Inc.,1992

84. Matthews DE, Farewell VT. Using and understanding medical statistics. (3rd. ed.) New York: Karger,1996

85. Talley NJ, Zinsmeister AR, Schleck CD, et al. Smoking, alcohol, and analgesics in dyspepsia and among dyspepsia subgroups: lack of an association in a community. Gut 1994;35:619-24.

86. Talley NJ, Weaver AL, Zinsmeister AR. Smoking, alcohol, and nonsteroidal anti-inflammatory drugs in outpatients with functional dyspepsia and among dyspepsia subgroups. Am J Gastroenterol 1994;89(4):524-8.

Appendix:

A) Statistical Formulae:

Calculation of the Intraclass Correlation Coefficient: (50)

1) Calculate the standard deviation for T1, T2, and their differences and label each subsequently as A, B, and C. Let D= the average difference.

2) Compute the Total Sum of Squares (S) for the ANOVA table.

$S=(n-1)(A^2+B^2) + nD^2/2$ where n= total number of samples

3) The "occasion" sum of squares = $nD^2/2$

4) The residual sum of squares = $(C^2/2)(n-1)$

5) The "person" sum of squares = $(A^2 + B^2 - C^2/2)(n-1)$

The values calculated above reproduce an ANOVA table.

$$ICC = \frac{MSP-MSE}{MSP + MSE(k-1) + 2(MSO-MSE)/n}$$

where MSP= Mean Square Person

MSO= Mean Square Occasion

MSE= Mean Square Error (Residual)

A simpler calculation for ICC would be

$$ICC = \frac{A^2 + B^2 - C^2}{A^2 + B^2 + D^2 - C^2/n}$$

if T1 and T2 were equal, then A=B, and C=D=0. The value of r = 1.

Calculation of Effect Size: (50)

Effect Size = (U-V)/E

where U = mean of group pre-treatment

V = mean of group post-treatment

E = standard deviation of group pre-treatment

Calculation of Guyatt Responsiveness Statistic: (50)

Guyatt Statistic = (U-V)/C

where U = mean of group pre-treatment

V = mean of group post-treatment

$C = \sqrt{2} \times MSE$

MSE is calculated by determining the Sum of Squares (residual) and dividing by n-1 degrees of freedom.

$SS\ (res) = (C^2/2)(n-1)$

where C = standard deviation of the differences between stable individual pre- and post-treatment.

Calculation of sample size based on the Guyatt Statistic:(36)

1) for independent groups

$2[(Z\alpha + Z\beta)\sigma/\Delta]^2$

2) for related groups

$[(Z\alpha + Z\beta)\sigma/\Delta]^2$

where α is the probability of Type 1 error (false positive rate) and β is the probability of Type 2 error (false negative rate or power) and σ the square root of the error variance $\sqrt{2}$ x MSE.

(see above for calculation of MSE.)

B) Tables:

Table 4.1: Nonulcer Dyspepsia Subgroups:

Dysmotility-like	upper abdominal pain assoc. with bloating nausea retching early satiety
Ulcer-like	upper abdominal pain which is localized to epigastrium often worse before eating relieved by eating or antacids awakens from sleep

Table 6.1: Manning Criteria for the Diagnosis of IBS.

pain decreased with defecation
looser stools with onset of pain
more frequent stools with onset of pain
visible abdominal distension
sense of incomplete evacuation
passage of mucus per rectum

Table A1: Stable Subjects before Treatment:			
Subject Number	Baseline P1	Followup P2	Difference P1-P2
1	20	20	0
2	12	9	3
3	9	12	- 3
4	9	6	3
5	15	15	0
6	16	16	0
8	15	15	0
10	8	9	- 1
11	16	12	4
12	9	9	0
13	25	25	0
14	20	12	8
15	12	9	3
16	15	16	- 1
17	9	9	0
18	6	9	- 3
19	15	15	0
20	20	20	0
21	20	12	8
23	10	10	0
24	6	6	0
26	25	20	5
27	20	16	4
28	6	9	- 3
30	15	15	0
31	9	12	- 3
32	12	12	0
33	25	25	0
34	15	12	3
35	6	9	- 3
37	15	25	- 10
38	9	9	0
39	6	10	- 4
40	20	20	0
41	12	12	0
42	6	9	- 3
43	20	20	0
44	6	9	- 3
Mean	13.5 (X)	13.4 (Y)	0.1 (D)
Standard Deviation	5.9 (A)	5.2 (B)	3.3 (C)
Variance	34.8	27.4	10.9

Table A2: Improved Subjects at Followup (After Treatment):			
Subject Number	Baseline P1	Followup P3	Difference P1-P3
1	20	1	19
3	9	4	5
5	15	6	9
6	16	6	10
9	6	1	5
10	8	4	4
12	9	1	8
13	25	20	5
14	20	4	16
17	9	9	0
19	15	1	14
20	20	4	16
22	9	9	0
24	6	1	5
28	6	6	0
29	9	6	3
30	15	15	0
32	12	6	6
33	25	1	24
34	15	9	6
38	9	4	5
39	6	1	5
40	20	15	5
41	12	4	8
42	6	6	0
43	20	6	14
44	6	6	0
Mean	12.9 (U)	5.8 (V)	7.1 (W)
Standard Deviation	5.8 (E)	4.7 (F)	6.4 (G)
Variance	33.7	22.4	40.6

Table A3: Subjects Who Improved A Little:				
Subject Number	Baseline P1	One Week P2	Followup P3	Difference P1-P3
13	25	25	20	5
17	9	9	9	0
23	10	10	10	0
24	6	6	1	5
28	6	9	6	0
29	9	9	6	3
30	15	15	15	0
34	15	12	9	6
40	20	20	15	5
Mean	12.8	12.8	10.1	2.7

C) Questionnaires:

QUESTIONNAIRE #1:

1) Have you been bothered on a regular basis by any of the following complaints? (Answer yes or no)

-abdominal pain _____

-nausea _____

-vomiting or retching _____

-upper abdominal bloating _____

-stomach fills up quickly when you eat _____

-other _____

2) Which one of these problems bothers you the most?

_____

3) In the past three months, how often have you had this problem? (Circle the best answer)

1) not at all.
2) once per week or less.
3) most days but not everyday.
4) once per day.
5) more than once each day

4) In the past three months, how severe has this problem been? (Circle the best answer)

1) No problem	
2) Mild Problem	can be ignored when you do not think about it.
3) Moderate problem	cannot be ignored but does not influence daily activities.
4) Severe problem	influences your concentration on daily activities.
5) Very severe	markedly influences your daily activities and/or requires rest.

5) How often have you used antacids in the past week? (Circle the best answer)

1) not at all.
2) once per week or less.
3) most days but not everyday.
4) once per day.
5) more than once each day

Patient Global Assessment:

Over the past week, how would you rate your stomach problem on the following scale?

(best it could be) (worst it could be)

Physician Global Assessment:

Rate the severity of the patient's symptoms (as you perceive).

none minimal mild moderate severe

QUESTIONNAIRE #2:

1) Have you been bothered on a regular basis by any of the following complaints? (Answer yes or no)

-abdominal pain _____

-nausea _____

-vomiting or retching _____

-upper abdominal bloating _____

-stomach fills up quickly when you eat _____

-other _____

2) Which one of these problems bothers you the most?

_____

3) In the past three months, how often have you had this problem? (Circle the best answer)

1) not at all.
2) once per week or less.
3) most days but not everyday.
4) once per day.
5) more than once each day

4) In the past three months, how severe has this problem been? (Circle the best answer)

1) No problem	
2) Mild Problem	can be ignored when you do not think about it.
3) Moderate problem	cannot be ignored but does not influence daily activities.
4) Severe problem	influences your concentration on daily activities.
5) Very severe	markedly influences your daily activities and/or requires rest.

5) How often have you used antacids in the past week? (Circle the best answer)

1) not at all.
2) once per week or less.
3) most days but not everyday.
4) once per day.
5) more than once each day

Patient Global Assessment:

Over the past week, how would you rate your stomach problem on the following scale?

(best it could be) (worst it could be)

QUESTIONNAIRE #3:

1) Since you were last seen, how often have you had your problem? (Circle the best answer)

1) not at all.
2) once per week or less.
3) most days but not everyday.
4) once per day.
5) more than once each day

2) Since you were last seen, how severe has this problem been? (Circle the best answer)

1) No problem

2) Mild Problem	can be ignored when you do not think about it.
3) Moderate problem	cannot be ignored but does not influence daily activities.
4) Severe problem	influences your concentration on daily activities.
5) Very severe	markedly influences your daily activities and/or requires rest.

3) How often have you used antacids in the past week? (Circle the best answer)

1) not at all.
2) once per week or less.
3) most days but not everyday.
4) once per day.
5) more than once each day

Patient Global Assessment:

Over the past week, how would you rate your stomach problem on the following scale?

(best it could be) (worst it could be)

4) Since you were last seen at the GI Unit one month ago, has there been any change in your stomach problem?
(Circle the best answer)

1) a great deal better.

2) moderately better.

3) a little better.

4) no change.

5) a little worse.

6) moderately worse.

7) a great deal worse.

Physician Global Assessment:

Rate the severity of the patient's symptoms (as you perceive).

none minimal mild moderate severe

Has there been any change in the patients stomach complaint?
(Circle the best answer)

1) a great deal better.

2) moderately better.

3) a little better.

4) no change.

5) a little worse.

6) moderately worse.

7) a great deal worse.

D) Nonulcer Dyspepsia Patient Subgroup Classification

Inclusion Criteria:

A) Ulcerlike subgroup:

Dyspepsia (recurrent upper abdominal pain > 3 months)
and 3 or more of the following:

	yes	no
pain/discomfort ac meal or when hungry		
night pain (waking from sleep)		
pain decreased with antacid		
periodic pain/discomfort		
well-localized pain/discomfort		

B) Dysmotility subgroup:

Dyspepsia (recurrent upper abdominal pain > 3 months)
and 3 or more of the following:

	yes	no
nausea at least once a month		
retching/vomiting at least once a month		
upper abdominal bloating		
abdominal pain worse with food/milk		
early satiety		
pain worse/discomfort worse after meals		

Exclusion Criteria:

- Documented organic disease at endoscopy.
- Prior gastric surgery.
- Pregnancy.
- Use in the past one month of

  Omeprazole

  antibiotics

  Pepto-Bismol
- Continuing use of NSAID.
- IBS patients (3 or more of the Manning criteria.)

  pain decreased with defecation

  looser stools with onset of pain.

  more frequent stools with onset of pain.

  visible abdominal distension.

  sense of incomplete evacuation.

  passage of mucus per rectum.

E) Data Sheets:

PATIENT DEMOGRAPHICS

Name:

Age:

Sex:

MCP#:

Home Address:

Phone #:

Medications:

Smoker: Amount smoked:

Alcohol consumption:

Caffeine consumption:

Concurrent medical problems:

Date seen:

Questionnaire # 2 to be filled out and returned on:

Follow-up visit on:

F) Computer software used:

Abascus Concepts Statview v.4.5 for the Macintosh

Claris Filemaker Pro v.3.0 for the Macintosh

Microsoft Excel v.4.0 for Apple Macintosh

Microsoft Word v.5.1 for the Macintosh

Niles and Assoc. EndNote Plus for the Macintosh

VITAE

NAME	Donald Garth MacIntosh
BIRTHDATE	Sept. 1, 1958
PLACE OF BIRTH	Montreal, Quebec
CITIZENSHIP	Canadian
HOME ADDRESS	18 Larch Place
	St. John's, NF.
	A1B 1R5
HOME TELEPHONE	(709) 726-4389
BUSINESS ADDRESS	Department of Medicine
	Health Sciences Centre
	300 Prince Philip Drive
	St. John's, NF.
	A1B 3V6
BUSINESS TELEPHONE	(709) 737-5070
BUSINESS FAX	(709) 737-3605
LANGUAGE	English
MARITAL STATUS	married, one child

CURRENT POSITION

Assistant Professor of Medicine (Gastroenterology)

Faculty of Medicine

Memorial University of Newfoundland

St. John's, NF.

Active Medical Staff

Division of Gastroenterology

The General Hospital Site

The Health Care Corporation of St. John's

St. John's, NF.

EDUCATIONAL BACKGROUND

1976-78	Undergraduate, University of Toronto
1978-79	Undergraduate, Memorial University
1979-83	Medical School, Memorial University
1983-84	Rotating Interne, Dalhousie University
1986-89	Medical Resident, Memorial University
1989-91	Fellow in Gastroenterology, University of Ottawa
1994-97	Postgraduate student in MSc. program in Clinical Epidemiology, Memorial University

Degrees and Certifications Awarded:

1981	B. Med. Sci. Memorial University
1983	MD. Memorial University
1984	LMCC
1989	Diplomat, American Board Internal Medicine
1990	FRCPC, Internal Medicine
1991	Certificate of Special Competence, Gastroenterology, Royal College of Physicians and Surgeons of Canada
1991	Diplomat, Subspecialty of Gastroenterology, American Board of Internal Medicine.

RESEARCH

Publications:

1) MacIntosh Donald G, Bear John C, Simpson John, et al. Should the Children of Patients with Hemochromatosis be Screened for the Disease? Can J Gastroenterol 1988. 2: 4: 143-46.

2) MacIntosh DG. and Leddin DJ. Transient Duodenal Ulceration in Association with Superior Mesenteric Ischemia. Can J Gastroenterol 1989, 3;1: 29-33.

3) MacIntosh DG. and Gillies RR. The Investigation of Dysphagia. Medicine North America 1991, 4; 17: 2294-2300.

4) MacIntosh DG, Thompson WG, Patel DG, et al. The Significance of Rectal Biopsy in IBS Patients. Am J Gastroenterol 1992, 87; 10: 1407-09.

Published Abstracts:

1) Fardy JM, Bursey F, and MacIntosh D. A Case Control Study of Smoking and Inflammatory Bowel Disease. Clin Invest Med 1995; 18(4): B48

2) Borgaonkar M, MacIntosh DG, and Fardy JM. A Meta-analysis of Anti-tuberculous Therapy for Crohns Disease. Can J Gastroenterol 1996; 10(Sup A): S35

3) MacIntosh DG, Fardy JM, Bursey F, and MacIntosh RF. Prevalence of Helicobacter pylori in Clinical Subgroups of Nonulcer Dyspepsia. Can J Gastroenterol 1997; 11(Sup A): S19

4) Borgaonkar M, MacIntosh DG, and Fardy JM. A Meta-analysis of Antibiotic Therapy for Crohns Disease. Can J Gastroenterol 1997; 11(Sup A): F6