Development and Validation of a Long Term Gout Response Criterion

Objective and Specific Aims:

A. Objective:

The primary objective of this study is to develop and validate a definition of a cessation or number and severity of gout flares, which will serve as an essential response criterion for therapy of chronic gout. In the past only non-validated clinical and laboratory outcomes have been used to assess the effect of interventions for gout. There is a lack of a valid clinically relevant response measure aimed at signs and symptoms for patients with chronic gouty arthritis.

We propose to define cessation or reduction in number and severity of gout flares in a specified period (e.g. 1 year) as the most important response criterion for treatments of chronic gout. Although a variety of domains have been identified as outcomes needing validation for chronic long term gout management all have limitations as described later. Appropriate treatments are likely to impact the flare criterion. Thus, the flare criterion will constitute a relevant and readily usable response criterion for comparison of treatments aimed at prevention in chronic gout.

B. Specific Aims:

The overall goal of this study is to develop a definition of elimination or limitation of number and severity of gout flares as a criterion for use in clinical trials and clinical practice. Gouty flare is a clinical outcome with construct/criterion validity but the definition has varied. A validated definition for use in trials and long term outcome studies is needed. There are two specific aims of this proposal:

1. To develop a definition of a gout flare and its severity which is necessary before we can use absence or number as a response criterion.
   Rationale: Gout flare is perhaps the most important outcome from both patients’ and physicians’ perspective. Although serum urate has been considered a surrogate marker for this very important outcome, the recent largest randomized trial of hypouricemic therapy in patients with chronic gout and hyperuricemia revealed that although the degree of change in uric acid differed between two therapies (febuxostat and allopurinol), the flare rates were similar (1). This suggests that there is a need to define and measure a change in or number of flares as a response criterion, rather than using surrogate definitions.

2. To use clinical trials to prospectively validate the definition of the gout flare criterion as a) a dichotomous response on presence vs. absence or b) tabulation of flare numbers, severity or a combination of the latter two as the response criterion.
   Rationale: For the response criterion to be applicable in gout prevention clinical trials, it must pass the Outcomes in Rheumatology Trials (OMERACT) Filter, i.e. be Truthful, Discriminative, and Feasible. This will be further detailed in the following pages. Sensitivity to clinical change i.e. responsiveness is one of the most critical components of any response criterion, since measures that lack responsiveness will have limited or no
use. A responsive and valid instrument will allow comparisons of efficacy of treatment strategies to each other or to a gold standard treatment.

C. Background and Preliminary Work:
C.1. Epidemiology and Scope of the Problem

Gout is a common medical problem that affects at least one percent of adult men in Western countries (2-4). It is the commonest cause of inflammatory arthritis in men older than 40 years of age (5). A gout flare signifies a significant increase in joint inflammation and symptoms. Gout flares are associated with functional disability, interference with work and recreational activities (6). In a recent cross-sectional study of 151 gout patients, patients with greater number of attacks or more severe attacks over the past year had lower scores on physical domains of the short-form 36 (J Hirsch, University of California, San Diego, personal communication). There are no prospective studies of frequency of gout flares or their health care impact; a major limitation is lack of a valid definition of gout flare. One of the goals of long term therapy of gout is to reduce the rate, severity and/or duration of gout flares. Development of a validated flare criterion definition will provide us with an effective response criterion to assess effectiveness of new therapies and to allow comparison between therapies.

C.2. Flare Criterion as Response Criterion for Patients with Chronic Gout

Patient-based outcomes are incorporated into response criteria used for various rheumatic diseases including rheumatoid arthritis, juvenile rheumatoid arthritis etc. (7-9). These include Stanford Health Assessment Questionnaire (HAQ) (10), modified HAQ (11) and children’s health questionnaire (CHQ) (12) as components of response criteria. There is currently no single validated laboratory or clinical variable that can measure disease activity or treatment response in patients with gout. Potential response criteria for chronic gout include pain, radiographic damage, tophus regression and health related quality of life (HRQOL) in addition to gout flares (13), but this proposal will focus on defining and validating the flare criterion for many reasons: (1) Flares are frequent in patients with gout and thus have the likelihood of being more responsive than other potential response criteria; (2) Pain and HRQOL may be affected by multiple comorbid conditions that most gout patients suffer from and it may be difficult to differentiate worsening associated with change in gout activity vs. change in activity of associated chronic conditions; (3) radiographic outcomes of joint destruction may take years to occur, and may not occur in all patients; (4) tophaceous gout occurs affects some but not all gout patients; and (5) definitions of disease flares have been developed for other rheumatic conditions such as lupus (14, 15) and ankylosing spondylitis (16) among others, and reduction of flares constitutes an important response measure in clinical studies of lupus. Cessation or number and severity of gout flares thus can constitute an important outcome criterion for patients with chronic gout and is high priority for both researchers and clinicians alike, particularly in light of new, emerging therapeutic options (17). An alternative approach we considered but rejected was to propose developing a composite outcome similar to ACR 20 for rheumatoid arthritis (18). However, none of the component measures such as radiographic changes, joint counts or tophus measurement have been validated. This would also be
considerably more complex and less feasible. Due to these limitations, we propose validating the flare criterion as the initial and hopefully definitive effort.

Flares are important in patients with chronic gout for the following reasons: (1) they are important clinical outcome for both patients and physicians; (2) they are frequent (19, 20); (3) they are associated with impaired functional status, work ability (6) and quality of life (J Hirsch, personal communication); and (4) effective use of standard therapies and/or newer therapies can decrease the frequency of flares. The frequency of gout flare has been reported in a wide range of two/year in untreated patients with chronic gout (21) to 16/year in those with refractory severe tophaceous gout despite optimal therapy (22). The frequency of gout flares is also influenced by treatment such that the flare frequency decreases (from two/year to 0.5/year) after initiation of colchicine prophylaxis (21) and the proportion of patients having flares temporarily increases (from 22-36% patients having flares during colchicine prophylaxis to 64-70%) when colchicine is withdrawn in patients receiving allopurinol or febuxostat. Long-term continuous use of allopurinol is associated with lower frequency of flares than intermittent use (23). Thus, flares are frequent in patients with chronic gout and effective treatment decreases the flare frequency.

To date, various studies have used different definitions for gout flare. These include: (1) Acute events requiring physical consultation (24); (2) acute attacks requiring medical intervention at physician’s discretion (1, 25); (3) acute events requiring non-steroidal anti-inflammatory drugs (NSAIDs) for short-term by physician (21); (4) acute events requiring NSAID or colchicine treatment (26); (5) joint symptoms of sufficient severity to require an emergency room or urgent outpatient medical attention (27); (6) Office or emergency room visit with claim for gout or joint pain (code) plus ≥1 of the following within 7 days post-visit (intraarticular aspiration or injection, joint fluid microscopy, or pharmacy claim for NSAID, colchicine, corticosteroid, or ACTH) (28). None of these definitions have been validated and each has its limitations. While patient report has the advantage of more inclusive approach to include all flares, mild and severe, it suffers from recall bias. Use of prescription anti-inflammatory treatment for gout flare has the potential advantage of being a more specific approach, but NSAIDs might also be used to treat pain other than that due to gout. In addition, it is likely to miss many milder flares that are self-limited, not treated with prescription medication, or treated with over-the-counter non-steroidal anti-inflammatory drugs (NSAIDs). Therefore, a validated definition of flare criterion is needed.

C.3. Preliminary Work Accomplished by our group
C.3.1. A New Preliminary Definition of Gout Flares
Recently a preliminary definition of gout flare was suggested by Grainger and colleagues using a Delphi consensus of rheumatologists (29). Of the 43 potential items 6 items were identified as “definitely appropriate for a gout flare definition (median score 7 to 9 with scale ranging from 1 (highly inappropriate) to 9 (highly appropriate)). These included; (1) Swelling of the affected joint; (2) redness of the affected joint; (3) marked tenderness of affected joint ; (4) urate crystals demonstrated from the joint; (5) History of gout; and (6) maximum pain in 4-12 hours. Of these, a prior history of gout or demonstration of urate crystals in joint fluid are not necessarily features of “gout flare”,
but rather similar to the classification criteria for gout. However certain other items such as acute onset, marked impairment of function with similarity to previous flares, pain preventing walking or a change of pain > 3 from baseline on a 10cm VAS were considered neither appropriate nor inappropriate by Delphi consensus for a gout flare definition. Some of these items may be relevant for defining the gout flare. These features need to be addressed in a further Delphi exercise, as proposed in this application. Severity of flares was not addressed but now will be.

As we will have need to identify flares not witnessed by physicians for use in some (or most) clinical trials the exercise will address whether, as is likely, some differences in criteria are needed for these flares compared to physician documented flares.

C.3.2. Outcomes Measurement in Rheumatology (OMERACT) Special Interest Group on Gout

OMERACT’s Special Interest Group in gout consists of international experts in the field of gout investigating various aspects of gout including classification and response criteria. OMERACT 7 proposed domains for assessment of both acute and chronic gout (Appendix A). Gout flare recurrence was prominent among the nine proposed core set domains for response in chronic gout (serum urate, flare recurrence, tophus regression, joint damage imaging, health related quality of life, musculoskeletal function, patient global assessment, participation and safety and tolerability) (13).

C.3.3. 1977 ARA Preliminary Classification Criteria for Acute Gouty Arthritis

Although widely used, these criteria concern diagnosis of gout, not just flares, address only acute gout and lack specificity. Six criteria as proposed have 87.6% sensitivity but 19.5% of other causes of arthritis surveyed would be misclassified as acute gout. Using 7 criteria could increase specificity but would decrease sensitivity (13). Therefore, response criteria for comparing treatments of chronic gout are needed.

D. Challenges with Definition of Gout Flare Criterion

One of the major challenges in defining flare criterion (cessation or number and severity of flares) is to define a gout flare. This is a challenging task due to a variety of reasons. Many patients may self-treat flares at home with NSAIDs or colchicine, and may not request prescription medications and/or visit health care professionals for their flare. Most experts would agree that this practice is quite prevalent and may be appropriate, in many cases. Thus, many flares may be witnessed only by the patients, and not health care professionals, but are still important. In a recent randomized controlled study of febuxostat vs. allopurinol for treatment of hyperuricemia in patients with gout, about 25% of gout flares reported by investigators did not receive any reported treatment. While requirement of treatment with medications such colchicine, prednisone or non-steroidal anti-inflammatory drugs signifies a definite occurrence of a gout flare in most cases, absence or lack of treatment may not equate absence of a flare, specially if the flare was mild. Some flares require visits to physicians in acute care or outpatient care setting and some may require additional therapy with prescription or non-prescription medications. These flares are likely to be more severe
and have impact on health care utilization and costs. Both types of flares (patient-reported and physician-examined) are likely to impact a patient’s quality of life, function and mobility. It is important to try to define both types of flares, although it is more challenging to validate candidate definitions for flares for which patients do not seek medical attention.

Severity of flares needs to be defined. One may define this as some magnitude of increased pain or some level of pain on a VAS. An alternative is to define flares categorically as mild, moderate or severe. Some may argue that use of over-the-counter medication, prescription medication, or health care visits for a flare may represent a hierarchy of flare severity. However, these behaviors may be influenced by other factors such as health care beliefs, health care access and insurance among others.

Duration of flare is an important aspect of flares. Effective therapies initiated early may abort flares or decrease the flare duration, while long-term therapies are expected to decrease flare numbers but also flare severity and duration. Thus, flare criterion definition will include frequency, severity and/or duration of flares. There will be continued discussion and investigation on how best to include severity and duration in the response criterion. Whether one needs to establish a baseline frequency of flares or aim for a set number (0-1) of flares per year as a part of flare criterion response needs to be considered.

It is likely that an absolute number of flares in a specified period of time (e.g. 1 year), severity of flares and their duration will be important to evaluate as for inclusion in our gout flare response criterion definition. In many clinical trials there can be comparisons with self-reported non-validated flares during the previous year and with validated flares during the first months of a treatment compared with the final months. Hopefully in some future trials there can be a pretreatment observation period of validated flares.

E. Research Design and Methods
E.1. Overview

The overall objective of this study is to define and validate the flare criterion as the response criterion for long term treatment of gout whereby the absence of or reduction in number and severity per specified period will be the desired outcome. This is important since gout is a common medical problem (30) with no validated response criteria. Gout flares are common in patients with intermittent chronic gout (19, 20), and associated with significant morbidity and reduction in health-related quality of life. A validated response criterion will allow precise measurement of effect of current incompletely evaluated and new anti-gout medications on disease outcomes in patients with gout.

Validation of the definition will be conducted using the OMERACT filter for outcome measures in rheumatology that defines truth, discrimination and feasibility as essential components (31, 32). After a preliminary identification of proposed domains and components of a gout flare starting with a detailed literature review, features on our recent Delphi exercise, and enriched with a survey of gout patients, consensus expert physician opinion of the evidence will be obtained with an additional Delphi approach.
followed by a small group Nominal Group Technique (NGT) meeting. This will help to identify key domains/questions to be included in the flare definition. Validity (construct and criterion), reliability and responsiveness to change will be tested by using the flare instrument in patients with and without expert opinion defined flares of varying severities and durations in 5-8 busy rheumatology clinics representing different countries and patient populations. Finally the new criteria will be used and validated in two to three planned prospective gout studies of new gout therapies (uricase and febuxostat planned in late 2006-2007). Elimination of flares will be compared with number, severity and combination of number and severity as to value as response criteria.

E.2. Timeline
The timeline for the collection of data is displayed in Table E.2. We propose to start the project in July and August 2006 by formulating questions. Our team includes experts in epidemiology and measurement, development and validation of response instruments that will help us successfully develop and validate this instrument in a three-year period.

Table E.2. Revised Timeline as funding was delayed. First installment from ACR Dec 2006

<table>
<thead>
<tr>
<th>Task</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature review &amp; Candidate definition development</td>
<td>September - December 2006</td>
</tr>
<tr>
<td>Delphi Expert opinion on items</td>
<td>February-April 2007</td>
</tr>
<tr>
<td>Revision of Items</td>
<td>May 2007</td>
</tr>
<tr>
<td>Face-to-face Nominal Group meeting</td>
<td>June 2007</td>
</tr>
<tr>
<td>Revision of Items and item selection and teleconference training of clinics staffs</td>
<td>July 2007</td>
</tr>
<tr>
<td>Testing the patient and physician exam criterion in clinic patients</td>
<td>August-September 2007</td>
</tr>
<tr>
<td>Validity, reliability and responsiveness testing of criterion in patients in clinical trials</td>
<td>December 2007-June 2009</td>
</tr>
<tr>
<td>Data analysis, manuscript preparation and submission</td>
<td>July-September 2009</td>
</tr>
</tbody>
</table>

E.3 Development of Definition of Gout Flare
E.3.1. Literature Review
An expert panel consisting of 3-5 gout experts will review the existing literature regarding gout flares. The group will synthesize the literature regarding different definitions of gout flares and elements used in these definitions into an evidence report. A review and discussion of validity data on flare definitions or components of flare definitions will be done. Based on the initial work from Taylor et al., we anticipate that potential domains will likely include (but will not be limited to) joint swelling, redness, pain, tenderness, limitation of joint range of motion, limitation of mobility or daily function, serological markers such as CRP or ESR and/or associated psychological/social aspects of flare. Scaling of individual items will be considered as appropriate.
E.3.2. Consensus Development with Delphi and Nominal Group Technique exercises

Well-recognized consensus techniques that are specifically designed to combine judgments from experts in an area will be used. These included the Delphi technique and Nominal Group Technique (NGT) (33, 34). The Delphi technique uses a series of well-defined questionnaire surveys. NGT is a structured face-to-face meeting of an expert panel designed to facilitate consensus formation on a topic. These techniques have been used extensively in development of outcomes measures in rheumatology including those for rheumatoid arthritis, juvenile rheumatoid arthritis, lupus etc.

In the first step, the Delphi technique, we will send two-sequential surveys to select and then rank variables used in routine practice to judge if a patient is having flare or not and to grade severity. The survey will be coordinated by Dr. Taylor and sent to 30-40 experienced practicing rheumatologists representing international gout experts and physicians interested in gout who will be invited to be a part of the gout focus group. These include physicians from Europe, USA, New Zealand, Asia and South America (Appendix B).

In the first survey clinicians will be asked to select up to 10 variables that they judge are important components of gout flare. Potential variables for flare components will include various types of variable, including those obtained by clinical history, examination or those that are treatment-related. In particular, this will include the domains of pain, inflammation, function, physical and emotional problems due to gout flare, patient and physician global assessment, activity limitation by gout flare, treatment complications, and other practical problems relevant to the patients. This exercise will define the minimum level of severity for acute gouty arthritis to be called a flare in this setting: history, use of treatment, visit to a doctor, physical examination with minimum features, etc. In the second survey, variables identified by at least 10 responders in the first questionnaire and additional variables from published trials of gout will be listed and sent to the panel. Physicians will be asked to select and rank the top 10 most important components of gout flare. Physicians will also be surveyed regarding the minimum and maximum number of core set components of gout flare definition and their willingness to use dichotomized index for clinical trials and clinical practice scenarios.

The second step of consensus will be the 1 to 2-day consensus conference of 5-8 clinicians based on results of the Delphi panel. This will consist of experts different from those who participated in the Delphi panel from 5-6 countries to do the NGT. All participants will receive the evidence report as well as the proposed core set of flare criterion from the Delphi exercise as above. Dr. Saag and Dr. Shewchuk, both with expertise in NGT, will lead the NGT session. For each step participants will be first asked to work individually and then express themselves in a guided discussion. The exercise will include the following steps: (1) Participants will be asked to classify all selected variables into specific domains of flare definition and additional domains may be proposed by individuals or Steering Committee members; (2) Participants than classify all the variables identified above into domains; (3) Participants will then select and rank the domains that should be included in the core set of flare definition and severity characterization; (4) Participants will then select the variables to be used for measurement of the selected domains; and (5) Participants will be asked to identify the
inclusion and exclusion criteria for trials in which the proposed core set for flare definition may be used. This will be held at an ACR meeting to meet our timetable or if timing changes at a EULAR site.

Delphi and NGT processes will lead to achievement of content (clarity and completeness) and face (clear and lack ambiguity or repetition) validity. Redundant and/or duplicative items will be excluded from the final core set of components of flare definition to attempt to simplify for feasibility of use.

E.4. Validation of Definition of Gout Flare Criterion

Assessment of its validity, reliability and responsiveness to change.

E.4.1. Validation Process Overview

Figure E.4.1. represents an overview of the validation process. Validity will evaluated in accordance with the OMERACT filter of truth, discrimination and feasibility. Truth encompasses the following aspects: content, face, criterion and construct validity (35). Discrimination includes the responsiveness to change (discriminant validity). Feasibility of a measure will be assessed by the ease of use and its operationalization. The final candidate definition will be chosen based on truth, discriminative ability and feasibility.

Figure E.4.1.  
Face and Content Validity  
Prospective Validation process

Face and content validity will be obtained by the consensus process that includes the Delphi and nominal group process for gout flares as described above.

E.4.2. Validation in 2 Settings: Study Population, Recruitment and Privacy Protection

We plan to prospectively validate the flare criterion definition in two settings, an international multi-center clinic validation and prospective randomized studies of new therapeutic agents, including febuxostat and pegylated uricase.
The first step will be assembly of multi-center clinic validation study (USA, Europe, New Zealand) that will prospectively collect data on 150-200 consecutive crystal-proven gout patients seen in the clinics of investigators’ gout clinics (20-30 gout patients per investigator/center). The physicians participating in this exercise will be different from those involved in deriving candidate definitions for the flare criterion (external validation). Patients will be excluded if they have pseudogout, other crystalline arthritis or other types of inflammatory arthritis. Patients and physicians will individually assess presence or absence of flare as well as severity (pain VAS; mild, moderate or severe; duration; need for medication) at clinic visits over a 3-month period. An independent trained nurse/research assistant will collect the data with regards to the potential components of flare definition at each site. Patients will be examined during and in-between flares to validate these criteria.

The second step will consist of prospective validation and comparison of the various flare criteria in randomized trials of new agents, including febuxostat and/or pegylated uricase, with the main focus on assessing responsiveness to change with therapy. We have preliminary approval/interest from TAP, Ipsen (in Europe) and Savient pharmaceuticals to allow incorporation of the flare definition in the data collection for their respective studies planned for late 2006 and 2007. It is our understanding that for clinical trials by TAP and Savient, the ACR classification criteria will be used to recruit patients. For the Ipsen trial, crystal identification will be used for patient recruitment. Other inclusion and exclusion criteria in the clinical trials will be determined by each pharmaceutical company. Self defined flares during the previous year will be required for each study.

We anticipate that in the multi-center international clinic validation study, patients with different demographics (age, gender and race), disease duration, different severity, tophaceous vs. non-topchaceous gout treated vs. non-treated will be included. Patients in the clinical trials may be more homogenous with regards to disease characteristics. However, the clinical trials of gout will likely include many academic centers and some community patients. Since these trials will recruit patients in both USA and Europe, this will provide us with fairly representative populations including minorities, subjects in various age strata, of both genders, and with varying duration and severity of gout. Since gout is more common in men than women, it is very likely that a higher proportion of our patients will be men. All patients participating in clinic validation study will be recruited after an approval has been obtained from local Human Studies committees, and will undergo standard informed consent process and all the data collected will be in accordance with the Health Information Privacy Protection (HIPPA) regulations.

There is no patient remuneration for the multi-center clinic flare criterion evaluation. Any compensation for the clinical trials will be decided by each pharmaceutical company. Free or subsidized parking and coverage of transportation costs on an individual basis will be provided to patients, as applicable and allowable.

E.4.3. Gender, Minority and Community patient Inclusion

Our aim is to have a representative sample of gout patients by including patients at clinics worldwide to have a representative sample from different ethnic groups.
E.4.4. Assessing Validity

Validity will be evaluated in accordance with the OMERACT filter of truth, discrimination and feasibility. Truth encompasses the following aspects: content, face, criterion and construct validity (35). Discrimination includes the responsiveness to change (discriminant validity). Feasibility of a measure will be assessed by the ease of use and its operationalization. The final candidate definition will be chosen based on truth, discriminative ability and feasibility.

E.4.4.1. Face and Content Validity

Content validity refers to comprehensiveness, or to how adequately the sampling of questions reflects the aims of the index. Face validity of a measure represents its clinical credibility. Face and content validity will be obtained by the consensus process as described above.

E.4.4.2. Criterion Validity

The criterion validity represents the agreement of the measure or its correlation with the “gold standard.” We will use both patient and physician global assessment as to their status as two surrogate gold standards. We will evaluate if patient global assessment correlates with physician global assessment and with flare response criterion that includes patient-witnessed flares. We expect that patient and physician global assessments will be highly correlated. As discussed previously (Section D), the inclusion of patient-reported flares in the flare response criterion has the potential to make the definition more sensitive and less specific than the definition that includes physician-reported or physician and patient-reported flares. A major challenge in including patient-only described flares is that they are more difficult to validate than physician-reported flares. For patients in the clinic validation cohort, we will assess the correlation of patient-witnessed gout flares with physician-witnessed flares by asking patients at each visit, if they are having a flare and if so, to grade the severity of flare. Depending on the correlation, response definitions including either or both types of flares will be tested.

Patients and physicians will assess a flare is present at the time of evaluation (yes/no) and if they have a flare, they will assess flare severity on a 0-10 VAS scale as well as whether mild, moderate or severe. A consensus will be obtained as to what is the minimum increase on 0-10 scale that would be considered clinically significant. By virtue of it being a clinical outcome, gout flare has criterion validity. It is the definition that must be validated. Candidate flare definitions based on percent change in various domains or core set variables will be tested (based on sections E.3.1. and E.3.2.). Although we propose cessation of flares as one of the proposed definitions of flare criterion, definitions that include flares as a continuous variable and composite definitions that incorporate the number of flares/year, severity of flares (mild, moderate or severe; or on ordinal scale of 1-3), and duration of flares will also be tested. Consensus process will generate all these candidate definitions. Additionally, definitions based solely on improvement in patient or physician global assessment will be tested as has been done previously by Giannini et al. (8). For the clinic validation study, we will study both flares for which patient started self-treating with over-the-counter medications and for which they had to seek medical attention.
For testing predictive ability, possible outcomes include disability, tophi and radiographic destruction. Due to short duration of the clinic validation study and clinical trials, we are limited in our ability to measure long term outcomes as listed above. Longer observational studies or clinical trials will be needed to assess predictive ability of the flare criterion. Radiographic destruction may be hard to assess especially if very few patients have radiographic destruction. An additional limitation is that radiographic destruction may be influenced by other factors such as duration and severity of hyperuricemia, which may not have a high correlation with flares. We anticipate that temporary functional limitation may be one of the components of the flare criterion, therefore disability assessment with loss of productivity may have high correlation with flare criterion and therefore not truly predictive.

E.4.4.3. Construct Validity

Construct validity will be tested using known groups validation and convergent validity. Since the severity of gout may vary by presence/absence of tophi, renal function, use of medications etc., one would expect a difference in flare rate based on these characteristics i.e. known groups validation. we will examine the flare criterion in patients in following categories, where we expect lower proportion meeting the flare criterion in the former vs. latter groups respectively: Tophaceous vs. non-tophaceous gout; treatment with stable dose of allopurinol vs. intermittent users; those getting anti-inflammatory prophylaxis with colchicine or NSAIDs at initiation of allopurinol vs. those not receiving the prophylaxis; those with longer duration of gout compared to those with shorter disease duration; and patients with higher vs. lower/normal levels of serum urate. A challenge with known groups validation is that patients are likely to have variety of combinations of the above characteristics and some change over time, especially for the patients seen in the clinics. Thus, it may be difficult to get “pure” populations based on these, and even despite controlling for other characteristics, we may be limited in our capacity in performing some of these comparisons due to a limited sample size.

Convergent validity will be tested by comparing performance of flare definitions to the global assessments, since we expect correlation between flares and patient and/or physician global assessment. Physical domains of HRQOL measures such as Health Assessment Questionnaire (HAQ) and Short-form 36 (SF-36) are likely to have similarity to the underlying construct of the flare criterion and are expected to have moderate correlation with it. For example, patients with frequent and/or severe flares are likely to have higher HAQ and lower SF-36 scores, denoting poor function and poor health status. One of more of these measures (HAQ, SF-36, physician and patient global) is/are expected to be incorporated in the prospective pharmaceutical clinical trials.

E.4.4.4. Reliability and Internal consistency:

Inter-observer reliability in the clinical patient cohort will be tested by having two physicians assess flares independently. Assessment of test-retest will be difficult since symptoms of flare may have abated in 7-14 day interval required for this testing, and the answers may have changed due to time lag. The internal consistency of various components of flare definition will be tested using Cronbach’s coefficient alpha (36). It
is likely that a flare criterion definition that includes various components of flare (number, severity and duration) will have lower internal consistency, but higher sensitivity to detect change as compared to a dichotomous definition (e.g. no flares in a year vs. one or more). As described in section E.5.3.4., we will consider slight/fair Cronbach’s alpha acceptable for internal consistency.

E.4.4.5. Responsiveness to Change and Discriminative Ability

In order to assess responsiveness, we will test the flare criterion definition in gout patients participating in the planned pharmaceutical clinical trials as noted above. Responsiveness to change will be determined by assessing the ability of each component of the flare definition to detect change and differentiation between study groups.

A clinically important change will correspond to each component’s ability to detect change when either the physician or patient global assessment as “gold standards” indicates a change. Both physicians and patients will grade their flare on a 0-10 VAS scale at the time of each flare evaluation. A consensus will be obtained as to what is the minimum increase on 0-10 scale that would be considered clinically significant.

Discriminative ability will be evaluated by the ability to distinguish patients who improved (based on global assessments) from those who did not improve or worsened. Physicians and patients will determine (independent of each other) whether disease had improved, was stable or had worsened, compared to baseline at the end of the clinical trial. Patients and physician will do this evaluation without the knowledge of their baseline global assessments or results of other core set variables that constitute the flare criterion.

E.4.4.6. Feasibility

Our objective is to develop a flare criterion definition that is easy to use in both clinical practice and clinical trial setting. It is possible that different flare definitions may emerge for practice vs. clinical trials depending on types of data (history by diary, physical examination, laboratory testing), frequency of data collection, and effort and costs associated with verification and interpretability. The purpose of the flare criterion definition is to assess the effects of long term treatment of chronic gout in prevention trials (how do I prevent or reduce the occurrence of gout signs and symptoms?). This will be an important outcome in efficacy trials comparing existing hypouricemic therapies to each other; comparing newer hypouricemic drugs (or anti-cytokine drugs) to existing therapies; comparing newer long term prophylactic anti-inflammatory medications against traditional prophylactic (e.g. colchicine); and comparing existing prophylactic therapies to each other. This criterion is not recommended for use in assessing treatment of acute gouty arthritis or gouty nephropathy.

Specifically in each of the 3 trials with which we plan to evaluate gout flares as the response criterion all components identified as important as components of flare presence and severity will be recorded as part of the trial outcomes. Exact timing of data collection will depend on times of contacts pre-specified in each trial. For example the primary outcome in the Febuxostat Allopurinol-Controlled European trial which will start in early 2007 by Ipsen, the primary objective is serum urate lowering to < 6 mg/dl
at 12 mo but all details of gout flare features occurring during the previous month will be obtained in interview with patients at 1, 2, 3, 4, 6, 8, 10 and 12 mo. Only flares occurring at times of visits will include detailed physician documentation. In the TAP proposed phase 4 study there will be 2 years of comparisons with febuxostat; allopurinol with flexible dosing and a placebo arm. Patient reported flares will lead to detailed interviews to document features, and patients seen during flares will have physician documentation.

E.5. Analyzes
E.5.1. Introduction
The primary objective of data analysis is to develop and test the gout-flare criterion for validity, reliability and responsiveness.

E.5.2. Specific Aim 1. To develop a definition of gout flare criterion.
The items will be tested for comprehensiveness and brevity. Components/items that show significant collinearity, as shown by Pearson’s correlation coefficient of ≥ 0.7 will be considered redundant and not included in the final core component set. The scoring of the individual items, domains and overall scale will be done as described in the sections E.3.2. and E.3.3. In addition to candidate definitions based on different percent changes in core set variables and based on global assessments alone, we will use generalized estimation equation (GEE) models to identify additional possible definitions of a flare (37). These models take into account multiple observations per patient and differential length of follow-up. The fit of the GEE models will be assessed by comparing chi-square values and scaled deviance (=deviance/degrees of freedom). This will help in selection of best components of flare criterion definition.

E.5.3. Specific Aim 2. To validate definition of gout flare criterion prospectively
E.5.3.1. Assessing validity
Testing of the criterion and convergent validity of our flare criterion will be accomplished by reference to a patient and/or physician defined “gold-standard.” Criterion validity of individual components or constructs of the flare criterion will be examined by testing the ability of each component to predict global assessment ratings using linear or multivariate logistic regression models, with each component as predictor and global assessment as outcome. A core set of flare components will be dichotomized according to the cut-offs obtained from the receiver operating characteristics (ROC) curve analysis (38). Determination of best cut-offs for each variable will help the physicians to determine whether a patient has improved based on change in that particular parameter.

For assessing the individual candidate flare criterion definitions, we will calculate sensitivity, specificity, positive and negative predictive values. We will perform ROC analyses for all candidate definitions. ROC curves are graph sensitivity on y-axis and (1-specificity) on x-axis. The best candidate definitions are located in the upper left quadrant of the graph and have area under the curve closer to 1. The best combination of candidate items separates flare from non-flare states. Logistic regression will help to
identify independent predictors of flare. ROC could help identify the right number of items to maximise sensitivity and specificity of the definition. Alternatively, different combinations could be applied to the dataset and the definition with the highest Chi-square value and the most acceptable face-validity is declared the winner. Also, we might want to define in advance, that any definition of improvement needs to be met by less than 20% of placebo treated patients.

For convergent validity (an aspect of construct validity) testing of definitions compared to global assessment of flare (gold standard), we will consider Spearman’s rank correlation of >0.7 as high, 0.4-0.7 as moderate and <0.4 as low. We hypothesize that the correlation of the flare criterion definition with the physician/patient gold standard will be moderate, thus avoiding redundancy.

E.5.3.2. Assessing reliability/reproducibility

Inter-rater reliability will be assessed by asking two independent physicians to assess flare in the multi-center clinic validation study. For measurements on a continuous scale, intra-class coefficient will be used to measure reliability. Intra-class coefficient (ICC) measures the average similarity of the subjects’ scoring on each of the two ratings, and ranges from –1 to +1 (39). As a special case of ICC, we will calculate for a kappa coefficient for ordinal or dichotomous data. Kappa coefficient calculates the extent of agreement beyond chance and ranges from below 0 to 1 (40). Kappa values below 0.2 indicates poor agreement, 0.21-0.4 fair, 0.41-0.6 moderate, 0.61-0.8 substantial and above 0.8 almost perfect agreement.

E.5.3.3. Assessing responsiveness and discriminative ability

In assessing responsiveness of an instrument, we will evaluate how well the gout flare criterion differentiates subjects who improve clinically as defined by patient and/or physician global assessment, and subjects who don’t improve in chronic gout clinical trial settings. The ability of each individual core set variable to detect clinically important change between baseline and end-of-study measurement (varies 1-3 years for the trials) will be measured using the standardized response mean (SRM). SRM is calculated as the absolute mean change in score divided by standard deviation of an individual’s change in score (41). Absolute values of 0.2, 0.5 or ≥0.8 on SRM are considered small, moderate or large effects, respectively (42, 43). In addition, we will also examine the responsiveness of each component of flare criterion to detect clinically important change as measured by minimal clinically important change on patient or physician assessment. Additional ways to measure responsiveness including the Receiver operating characteristic (ROC) curve, a plot of sensitivity against 1-specificity will be used (44). Area under the curve provides an indicator of usefulness of candidate definitions and flare criterion definitions with an AUC >0.5 are considered responsive to change, as described previously (45).

Discriminative ability will be assessed by comparing patients who improved compared to those judged to have not improved (stable or worsened disease) by global assessments by t-test or the Mann-Whitney U test, as appropriate. We will also examine the level of agreement between physicians and patients in the evaluation of the response to therapy with a kappa statistic. If there are conflicting results based on physician vs. patient’s assessment, a consensus process will be performed.
E.5.3.4. Assessing internal consistency

Internal consistency between various items and domains, domains and overall score will be assessed by the Cronbach’s coefficient alpha. The cut-offs are as follows: 
<0.6 = poor, 0.6-0.64 = slight, 0.65-0.69 = fair, 0.7-0.79 = moderate, 0.8-0.89 = substantial, and >0.9 = almost perfect. We will consider that a slight/fair Cronbach’s alpha acceptable for internal consistency of the core set of flare components.

E.6. Quality Control

Consistent and accurate data collection is a top priority for our project. In our study extensive efforts will be made to assure consistency of measurements. Skilled study coordinators trained together by us will assist in data collection. Experienced clinicians will conduct the patient interview with examination. The physicians selected for definition development and validation of flare criterion will be different to avoid bias. The flare criterion will be tested in both clinic validation study and 2-3 clinical trials as proposed to allow for inclusion of different patient populations.

E.7. Data Processing

E.7.1. Data Entry

Standard data entry precautions such as random checking, double data entry, comparison for accuracy will be taken. Discrepancies will be resolved and errors will be corrected. Whenever errors are identified, an error report will be submitted to the data coordinator for review.

E.7.2. Storage and Analysis Systems

The Rheumatology division at the Philadelphia VA Medical Center has a research office close to the Rheumatology clinic. The research office is equipped with a computer that is used only for managing research data. In addition, there are locked file cabinets, where all confidential research data is stored. All patient files and records will be kept in the locked up cabinets and the study coordinator will have the key. A similar facility is available at Minneapolis VA Medical Center. Data entry will be done on the research computer in the Rheumatology research office. The data files will be password protected. In addition, a password protected back-up copy of the data will be kept on a second hard drive. Only the study coordinator, the PI and the co-investigators involved in data collection and analysis will have the password to access the data.

E.8. Organization and Personnel

E.8.1. Faculty

Key personnel for the project are described in Table E.8.1. Members of OMERACT gout group, ACR and EULAR will be investigators participating in this study.
(Appendix A). Dr. Ralph Schumacher will be the Principal Investigator. Steering committee/advisory board will consist of Drs Schumacher, Boers, Saag, Singh, Taylor and Perez-Ruiz. Dr. Taylor will lead the Delphi. Drs. Saag and Shewchuk will lead the NGT session. Regular meeting of the PI, study investigators, advisory board/steering committee and coordinators will occur through conference calls every 3 months and face-to-face during the annual ACR and EULAR meetings.

Table E.8.1. Gout Flare Definition Research Team

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ralph Schumacher, MD</td>
<td>Principal Investigator, University of Pennsylvania</td>
</tr>
<tr>
<td>Jasvinder Singh, MD, MPH</td>
<td>Co-Principal Investigator, Minneapolis VA Med Center</td>
</tr>
<tr>
<td>Kenneth Saag, MD, MPH</td>
<td>Co-investigator, University of Alabama</td>
</tr>
<tr>
<td>Fernando Perez-Ruiz, MD</td>
<td>Co-investigator, Hospital de Cruces, Baracaldo, Spain</td>
</tr>
<tr>
<td>Michael Doherty, MD</td>
<td>Co-investigator, Nottingham University, United Kingdom</td>
</tr>
<tr>
<td>Weiya Zhang, MD</td>
<td>Co-investigator, Nottingham University, United Kingdom</td>
</tr>
<tr>
<td>Maarten Boers, MD</td>
<td>Co-investigator, VU University Medical Center</td>
</tr>
<tr>
<td>William Taylor, MD</td>
<td>Co-investigator, University of Wellington, New Zealand</td>
</tr>
<tr>
<td>Lan Chen, MD, PhD</td>
<td>Co-investigator, University of Pennsylvania</td>
</tr>
<tr>
<td>Vibeke Strand, MD</td>
<td>Co-investigator, Stanford University School of Medicine</td>
</tr>
<tr>
<td>Maren Mahowald, MD</td>
<td>Co-investigator, University of Minnesota</td>
</tr>
<tr>
<td>Eswar Krishnan, MD</td>
<td>Co-investigator, University of Pittsburgh</td>
</tr>
<tr>
<td>Thomas Bardin, MD</td>
<td>Co-investigator, Hospital Lariboisiere, Paris, France</td>
</tr>
<tr>
<td>Richard M Shewchuk</td>
<td>Nominal Group Meeting leader, University of Alabama</td>
</tr>
<tr>
<td>Thomas Ten Have, PhD, MPH</td>
<td>Statistician, University of Pennsylvania</td>
</tr>
</tbody>
</table>

E.8.2. Project Personnel

The project personnel will consist of part-time study coordinators at various participating centers supervised by the local PIs. We will have a meeting to train and standardize the data collection for the nurse/research assistants from each center participating in the clinic validation study.

E.9. Strengths and Limitations

E.9.1. Recruitment & Generalizability

The OMERACT gout special interest group consists of about 30 Investigators and assuming a participation of 5-8 investigators of the group with 20-30 patients per investigator, we will have 150-200 patients in the clinic validation aspect study. If for some reason we have a lower than expected recruitment, we will expand our study to invite additional ACR and EULAR members with interest in gout to participate.

It is likely the many patients following up with investigators at tertiary medical centers have more severe form of gout or have a higher likelihood to receive gout treatment, due to referral bias and investigator’s interest, and may not be truly representative of community patients. In addition, most participating investigators are rheumatologists with interest in gout, and their disease assessment and management practices may differ from primary care physicians. However, inclusion of primary care physicians in this first phase of validation although desirable, is not practical. We expect that the subsequent validation of this flare definition in the prospective
pharmaceutical trials will occur in all clinic settings including primary care and family practice, since many gout patients are managed in these settings. Clinical trial participants may come from both primary care and rheumatology practices, but may differ from the general population in being more motivated, healthier, and in having lower comorbidity.

E.9.2. Ambiguous Items
It is possible that despite our best efforts some items in the questionnaire may be ambiguous or not comprehensive. Our multi-step systematic approach to the process, inclusion of Drs. Saag and Boers both of whom have experience in developing outcome instruments, formation of a Steering Committee and regular meetings with them are all steps designed to avoid this problem. If at any step, a question is raised about clarity or comprehensiveness of an item, this will be referred to the Steering Committee for resolution.

E.9.3. Limited Validity Testing
Due to absence of any gold standard, criterion validity of our questionnaire cannot be definitively tested. This problem is not unique to gout, but is shared by many other diseases. Realizing this limitation, we will compare our flare definition to physician and/or patient global questionnaire. In addition, validity testing will be more limited for testing patient reported flares as compared to the physician examined flares. The specialist opinion has served as a reasonable gold standard in many settings and against similar items on SF-36 or HAQ. Collection of flare criterion components by research assistants independent of the physicians’ or patients’ global assessment will avoid bias. Prospective testing in 2-3 different clinical trials will add to validity and responsiveness testing.

E.9.4. Responsiveness to change
We propose in this study that responsiveness to change will be tested in patients in clinical trials. It is possible that an instrument responsive to change in gout patients in clinical trial may not be responsive in clinical practice setting.
References:
12. Landgraf JM, Abetz L, Ware JE. The CHQ user’s manual. 1st ed. Boston: The Health Institute, New England Medical Center; 1996.


Appendices

Appendix A J Rheum 2005; 32:2452-2455. OMERACT 7 SIG: Outcome Measures for Acute and Chronic Gout
Appendix B OMERACT Gout Participants
Appendix C Letters of Support
Appendix D Biosketch of Statistician