**Abstract**

**Background:** Previous research has highlighted discrepancies between patient and physician ratings of disease severity. Whether this is also the case in psoriasis requires investigation. We investigate using study data on tofacitinib, an oral Janus kinase inhibitor that is being investigated for psoriasis.

**Objective:** To examine the level of concordance between two measures—Patient Global Assessment (PtGA) and Physician Global Assessment (PGA)—at baseline and at primary endpoints.

**Methods:** Pooled data were analyzed from four Phase 3 clinical trials (3,641 patients) of oral tofacitinib for the treatment of patients with moderate-to-severe chronic plaque psoriasis. Descriptive statistics and weighted kappa statistics were calculated. A repeated measures model was applied with PtGA as outcome and PGA as predictor.

**Results:** At baseline, physicians tended to rate the disease as “moderate” on the PGA while patients rated it as “severe” on the PtGA (weighted kappa=0.05). At post-baseline, there was moderate-to-substantial level of agreement (weight kappa=0.61). At baseline correlations were consistently low across the four trials (0.09, p<0.01 in all cases), whereas by the primary time point the correlations were substantially higher for each trial (range: 0.65–0.71, p<0.001 in all cases), as a result of trial inclusion and exclusion criteria. In general, a clear trend surfaced with lower (less severe) PGA scores systematically corresponding to higher (more severe) PtGA scores. For example, for a PGA score of 0 (“clear”) the corresponding PtGA score was 0.987, numerically very close to 1 (“almost clear”).

**Conclusion:** In psoriasis the physician ratings of disease severity tend to consistently rate severity lower than the corresponding patient ratings of severity.

**ABBREVIATIONS**

BID: twice daily; BIW: twice weekly; FDA: Federal Drug Administration; PGA: Physician Global Assessment; PtGA: Patient Global Assessment

**INTRODUCTION**

Assessing the severity of disease over time is important for optimizing patient care, since it allows for the critical evaluation of individual response to treatment. In addition, assessing disease severity is often a key aspect of quality care and quality guidelines, which are increasingly important to improve the level of care in the community.

Assessing the severity of plaque psoriasis can be difficult because of the chronic nature of the disease and the variable disease impact on patients. There are many different ways of assessing psoriasis severity, although no one measure encompasses all aspects of the disease. In clinical trials, the...
Physician Global Assessment (PGA) is recommended by the FDA as an efficacy end point, with its summary score used to measure overall disease severity. The PGA requires physicians to evaluate the severity of plaques for three clinical signs (erythema, induration, and scaling) across the whole body. Although the PGA does not take into account other aspects of psoriasis, such as amount of body surface area affected, pain or pruritus, it is routinely used in clinical trials since it provides a relatively simple assessment of overall disease severity.

In the Phase 3 clinical trial program for tofacitinib, a Janus kinase inhibitor currently in development for the treatment of moderate-to-severe plaque psoriasis, a 5-category PGA was used as a co-primary endpoint [1]. As a key secondary endpoint, the Patient Global Assessment (PtGA), a patient-reported global measure of disease severity, was included in this clinical program. The PtGA used the same 5 category labels as the PGA for response options (i.e. “clear,” “almost clear,” “mild,” “moderate and “severe”). However, unlike the PGA, the PtGA did not include descriptors for the clinical signs of erythema, induration, and scaling to guide the assessment. Instead, patients were simply asked to rate the overall severity of their psoriasis at the current time. Increasingly, disease severity assessment in chronic diseases should be a joint evaluation by patient and physician [2-4].

The PtGA was designed to be a patient-based assessment of overall disease severity to complement the clinician-based assessment provided by the PGA. As the PtGA is the patient’s interpretation of their overall disease status without necessarily focusing on the plaque. It is therefore likely that when answering the PtGA, patients may have considered other aspects of the disease that are important to them and that extended beyond erythema, induration, and scaling. Patient-oriented research has shown that the most bothersome features of psoriasis to patients were itching, scales and flaking [5], symptoms which are not directly captured by the PGA.

Previous research has highlighted discrepancies between patient and physician ratings of disease severity, with physicians tending to underestimate or underreport symptoms compared to patients in various diseases [6-8]. We were interested in examining whether this is also the case in a visible disease such as psoriasis. To that end, we examined the level of concordance between the PtGA and PGA at baseline and at the primary time point of the tofacitinib Phase 3 clinical trial program.

METHODS

Patient and studies

Pooled data from four Phase 3 clinical trials (total n = 3641) of oral tofacitinib for the treatment of patients with moderate-to-severe chronic plaque psoriasis were analyzed. Full details of the 4 trials (A3921078, NCT01276639; A3921079, NCT01309737; A3921080, NCT01241591; and A3921111, NCT01186744) are described elsewhere [9-11]. Briefly, patients were randomized to receive tofacitinib 5 mg BD or tofacitinib 10 mg BD; or placebo (in 3 trials) and one trial included an active control (A3921080, etanercept 50 mg/BIW). Patients had to have a score of “mild” or “severe”, assessed by an investigator, on the PGA scale at baseline for inclusion into the trials. Across the four trials, 27 patients with a “mild” PGA score at baseline were randomized and included in the trials; these were protocol deviations but were included in the analyses. The PGA was assessed at every clinical trial study visit by a trained investigator. Whenever possible, the same investigator assessed all the efficacy endpoints for an individual patient throughout the study. The PtGA was self-reported by patients and assessed at each clinic visit like the PGA. The recall period for the PGA and PtGA was at the current point in time.

Statistical analyses

Descriptive analyses that included counts, percentages, and Pearson correlation coefficients of the two variables (PGA and PtGA) were obtained. In addition, as a measure of agreement between the two variables above and beyond chance, the weighted kappa statistics were calculated [12,13]. Kappa statistics can be interpreted as follows: ≤0, poor; 0–0.2, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect [14]. For the correlations and frequency tables, two time points were examined: baseline and post-baseline, with post-baseline being the primary time point for each of the four studies [week 16 (studies A3921078 and A3921079), week 12 (A3921080) and the end of the does-blind active treatment period (week 24 in study A3921111)].

To study the relationship between PGA and PtGA, we employed a repeated measures model with PtGA as the outcome and PGA treated as a continuous predictor (anchor)(which means that we imposed a linear relationship between the outcome and the predictor). All available data were used (including baseline and all post-baseline clinical visits) from studies A3921078, A3921079, A3921080, but from study A3921111 only data from the first 24 weeks when all subjects received active treatment (which involved only its randomized portion) were used. The different design of every study was accounted for by using the different variance-covariance matrix (by using the GROUP=Study option in the MIXED procedure in SAS).

As a sensitivity analysis to assess the linearity assumption, PGA score was used as a categorical predictor. Using PGA as a categorical anchor does not impose any functional relationship between it and the PtGA.

All analyses were performed in SAS version 9.4 [15].

RESULTS AND DISCUSSION

The two-way contingency table analysis of PGA-by-PtGA at baseline showed that for 52.7% of the observations physicians rated the disease as “mild” on the PGA while patients rated it as “severe” on the PtGA (Table 1). The weighted kappa statistic was 0.05 at baseline, indicating a poor level of agreement. In contrast, the post-baseline weighted kappa was 0.61, indicating a moderate-to-substantial level of agreement (Table 2).

The Pearson correlation coefficients are shown in (Table 3). At baseline, the Pearson correlation coefficients were consistently low across the four trials (0.09, p<0.01 in all cases), whereas by the primary time point the Pearson correlation coefficients were substantially higher for each trial (range: 0.65–0.71, p<0.001 in all cases).
Our analyses based on data from four Phase 3 trials of tofacitinib show that there are indeed differences in how patients and dermatologists assess overall psoriasis severity, with patients consistently tending to rate their psoriasis as more severe than physicians. Both frequency tables at baseline and post-baseline show the same pattern of results, but with very different kappa statistics: for any given PGA score (except a score of 4), PtGA scores were shifted to higher (i.e., more severe) values. Note that these results, which were taken from just two time points, correspond exactly to the results from the functional relationship, which used all available data (Figure 1). These results indicate that patients consistently rated their psoriasis as more severe than their physician. Although physicians were constrained by the clinical descriptors when assessing severity and the scoring algorithm of the PGA, it remains noteworthy that the patient interpretation of their overall disease status tends to be more severe.

At post-baseline the weighted kappa statistic is 0.61, suggesting “substantial” agreement between PGA and PtGA. Although the weighted kappa statistic was small (0.05) at baseline, this does not mean that there is no relationship between the two variables. As previously described, at baseline there is the same pattern of relationship between PGA and PtGA as that observed post-baseline. At baseline, however, the response options for the PGA are limited to “moderate” or “severe” as a result of the trial inclusion and exclusion criteria, which restricted patients to have a similar set of responses before treatment intervention. However, the same trend was observed post-baseline when the full range of responses on the PGA was possible, owing to the effects of the interventions. As expected, post-baseline observations provide a much broader range of responses (attributed to the beneficial effects of tofacitinib) for both PGA and PtGA. Such a larger variation of scores increases the correlation and agreement between the two measures.

CONCLUSION
The PGA is a clinical measure assessing the erythema, induration, and scaling of psoriatic plaques across the entire body in order to give an overall profile of disease severity. The PtGA, in contrast, is a subjective assessment by the patient of their overall disease severity, which may or may not take into account erythema, induration, and scaling, as well as other disease features such as pruritus or the painfulness of plaques. In spite of the different natures of these endpoints, the fact that both assess disease severity lends itself to interesting comparisons between how physicians and patients view overall psoriasis severity. Our analyses indicate that in psoriasis, the physician ratings of disease severity tend to consistently rate severity lower than

---

**Table 1:** Contingency table of the pooled data at Baseline showing the frequency of PGA and PtGA responses by category.

<table>
<thead>
<tr>
<th>N (%)</th>
<th>PtGA 0 (clear)</th>
<th>PtGA 1 (almost clear)</th>
<th>PtGA 2 (mild)</th>
<th>PtGA 3 (moderate)</th>
<th>PtGA 4 (severe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGA 0 (clear)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PGA 1 (almost clear)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PGA 2 (mild)</td>
<td>0</td>
<td>0</td>
<td>2 (0.06)</td>
<td>11 (0.30)</td>
<td>14 (0.39)</td>
</tr>
<tr>
<td>PGA 3 (moderate)</td>
<td>0</td>
<td>4 (0.11)</td>
<td>97 (2.69)</td>
<td>1046 (28.97)</td>
<td>1904 (52.73)</td>
</tr>
<tr>
<td>PGA 4 (severe)</td>
<td>0</td>
<td>1 (0.03)</td>
<td>10 (0.28)</td>
<td>127 (3.52)</td>
<td>395 (10.94)</td>
</tr>
</tbody>
</table>

**Abbreviations:** PGA: Physician Global Assessment; PtGA: Patient Global Assessment

**Table 2:** Contingency table of the pooled data Post-Baseline showing the frequency of PGA and PtGA responses by category.

<table>
<thead>
<tr>
<th>N (%)</th>
<th>PtGA 0 (clear)</th>
<th>PtGA 1 (almost clear)</th>
<th>PtGA 2 (mild)</th>
<th>PtGA 3 (moderate)</th>
<th>PtGA 4 (severe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGA 0 (clear)</td>
<td>232 (7.27)</td>
<td>275 (8.62)</td>
<td>43 (1.35)</td>
<td>17 (0.53)</td>
<td>5 (0.16)</td>
</tr>
<tr>
<td>PGA 1 (almost clear)</td>
<td>48 (1.50)</td>
<td>565 (17.71)</td>
<td>333 (10.44)</td>
<td>162 (5.08)</td>
<td>24 (0.75)</td>
</tr>
<tr>
<td>PGA 2 (mild)</td>
<td>6 (0.19)</td>
<td>153 (4.79)</td>
<td>259 (8.12)</td>
<td>315 (9.87)</td>
<td>136 (4.26)</td>
</tr>
<tr>
<td>PGA 3 (moderate)</td>
<td>0</td>
<td>22 (0.69)</td>
<td>82 (2.57)</td>
<td>211 (6.61)</td>
<td>228 (7.15)</td>
</tr>
<tr>
<td>PGA 4 (severe)</td>
<td>0</td>
<td>3 (0.09)</td>
<td>8 (0.25)</td>
<td>21 (0.66)</td>
<td>43 (1.35)</td>
</tr>
</tbody>
</table>

**Abbreviations:** PGA: Physician Global Assessment; PtGA: Patient Global Assessment

**Table 3:** Pearson correlation coefficients (p-values) between PGA and PtGA for each trial.

<table>
<thead>
<tr>
<th>N (%)</th>
<th>A3921078</th>
<th>A3921079</th>
<th>A3921080</th>
<th>A3921111</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.095 (0.005)</td>
<td>0.093 (0.004)</td>
<td>0.096 (0.001)</td>
<td>0.096 (0.013)</td>
</tr>
<tr>
<td>Post-baseline</td>
<td>0.704 (&lt;0.0001)</td>
<td>0.71 (&lt;0.0001)</td>
<td>0.660 (&lt;0.0001)</td>
<td>0.701 (&lt;0.0001)</td>
</tr>
</tbody>
</table>

**Abbreviations:** PGA: Physician Global Assessment; PtGA: Patient Global Assessment

Figure 1 represents the PtGA score for every category of the PGA, with PGA as a continuous anchor. A clear trend was observed for PGA scores of 0, 1, 2, and 3: the lower PGA scores systematically corresponded to higher (i.e., more severe) PtGA scores. As depicted in Figure 1, for example, for a PGA score of 0 (“clear”) the corresponding PtGA score is 0.987, which is numerically very close to 1 (“almost clear”) (Figure 1). The same kind of pattern was observed for PGA scores of 1 (“almost clear”), 2 (“mild”) and 3 (“moderate”). At the high end of the scale (PGA score of 4) both scales converge due to the scale designs (the upper limit of the PGA and PtGA range).

At post-baseline the weighted kappa statistic is 0.61, suggesting “substantial” agreement between PGA and PtGA. Although the weighted kappa statistic was small (0.05) at baseline, this does not mean that there is no relationship between the two variables. As previously described, at baseline there is the same pattern of relationship between PGA and PtGA as that observed post-baseline. At baseline, however, the response options for the PGA are limited to “moderate” or “severe” as a result of the trial inclusion and exclusion criteria, which restricted patients to have a similar set of responses before treatment intervention. However, the same trend was observed post-baseline when the full range of responses on the PGA was possible, owing to the effects of the interventions. As expected, post-baseline observations provide a much broader range of responses (attributed to the beneficial effects of tofacitinib) for both PGA and PtGA. Such a larger variation of scores increases the correlation and agreement between the two measures.
the corresponding patient ratings of severity. Future work could examine whether this relationship holds true for other endpoints routinely used to assess psoriasis, such as the Psoriasis Area and Severity Index and the Dermatology Life Quality Index.

Although the PGA is not routinely used in clinical practice, these analyses highlight that there may indeed be a different perception of disease severity between patients and their physicians, even for a highly visible disease such as psoriasis. One limitation of this study is that there is no standard PGA scale that is used across trials. Thus these results are based on one version of a PGA scale and findings may be different with different versions of the PGA. However we believe this is unlikely, because similar findings have been reported for a wide range of other diseases, such as major depressive disorder [16], asthma [17], gastroesophageal reflux disease [18], gout [19], and AIDS [20], with physician-report typically underestimating the severity of symptoms compared to patient-report. It is therefore important to take into account the patient perspective in treatment decisions. Current treatment goals and guidelines have shortcomings as they rarely include the patient perspective on treatment effectiveness. Although the PGA and PtGA can be considered complementary assessments of global disease severity, they are not exhaustive measures and as such it may be useful to also assess other aspects of importance to patients, such as itch, pain, and quality of life. The International Dermatology Outcome Measures (IDEOM) group is actively working to address this issue in psoriasis by establishing outcome measures of relevance to clinicians and patients [21]. Future recommendations need to include the patient perspective for a better match between treatment decisions and patient expectations.

DISCLOSURE

This publication reports the results of a Pfizer-sponsored study.

CONFLICT OF INTEREST

AB, JC, LM, and CM are Pfizer employees and own Pfizer stock.

CP has been a consultant or investigator for the following companies: Abbvie, Amgen, Bohringer, Celgene, Eli-Lilly, Janssen-Cilag, Novartis, and Pfizer.

REFERENCES


9. NCT01276639 (A3921078) and NCT01309737 (A3921079).


15. SAS Institute Inc.


