Title
Data-driven desirability function to measure patientsâ�� disease progression in a longitudinal study

Journal
Journal of Applied Statistics, 43(5)

ISSN
0266-4763 1360-0532

Authors
Chen, Hsiu-Wen
Wong, Weng Kee
Xu, Hongquan

Publication Date
2015-10-09

DOI
10.1080/02664763.2015.1077378

Peer reviewed

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Data-driven desirability function to measure patients' disease progression in a longitudinal study

Hsiu-Wen Chen, Weng Kee Wong & Hongquan Xu

To cite this article: Hsiu-Wen Chen, Weng Kee Wong & Hongquan Xu (2015): Data-driven desirability function to measure patients' disease progression in a longitudinal study, Journal of Applied Statistics, DOI: 10.1080/02664763.2015.1077378

To link to this article: http://dx.doi.org/10.1080/02664763.2015.1077378

Published online: 09 Oct 2015.

Journal of Applied Statistics, 2015
http://dx.doi.org/10.1080/02664763.2015.1077378

Taylor & Francis
Taylor & Francis Group

Data-driven desirability function to measure patients' disease progression in a longitudinal study

Hsiu-Wen Chen[a*], Weng Kee Wong[b] and Hongquan Xu[c]

[a]*Department of Industrial and Systems Engineering, Chung Yuan Christian University, Taoyuan City 32023, Taiwan;* [b]*Department of Biostatistics, University of California, Los Angeles, CA 90095, USA;* [c]*Department of Statistics, University of California, Los Angeles, CA 90095, USA*

(*Received 14 February 2015; accepted 25 July 2015*)

Multiple outcomes are increasingly used to assess chronic disease progression. We discuss and show how desirability functions can be used to assess a patient overall response to a treatment using multiple outcome measures and each of them may contribute unequally to the final assessment. Because judgments on disease progression and the relative contribution of each outcome can be subjective, we propose a data-driven approach to minimize the biases by using desirability functions with estimated shapes and weights based on a given gold standard. Our method provides each patient with a meaningful overall progression score that facilitates comparison and clinical interpretation. We also extend the methodology in a novel way to monitor patients' disease progression when there are multiple time points and illustrate our method using a longitudinal data set from a randomized two-arm clinical trial for scleroderma patients.

Keywords: desirability function; longitudinal data; multiple outcomes; nonlinear least squares; scleroderma

1. Introduction

Desirability functions have enjoyed sustaining popularity in the industrial sector but their use in the medical sciences has been rather limited to date. Their origin and use were motivated by the need to use several outcome measures to evaluate the overall quality of a product using a single score. Each component of the composite measure is given a desirability score and the overall desirability score is obtained by combining the individual scores from all components in the composite measure. High scores represent better quality and they are conveniently scaled between 0 and 1, with 1 representing the best possible product. Desirability function was first

*Corresponding author. Email: hwchen@cycu.edu.tw

proposed by Harrington [10] and remains a highly active area of research in non-biomedical fields. Some examples include use of such functions to interpret a composite outcome when there is model uncertainty [11] or there is heteroscedasticity in the component outcomes [9].

Clearly, it is challenging to combine multiple outcomes into an overall composite score with a meaningful interpretation. When there is only a single score, there are usually established standards for interpreting the score in a meaningful way. When there are several outcomes giving different scores, it is less clear how to combine them and interpret the scores taken as a whole. There are several difficulties involved. First, the outcome measures may vary in conflicting directions or there is dependence among these outcomes. Second, the outcomes cannot be directly combined because they are measured in different units. Third, multiple outcomes are likely to have different levels of importance in terms of their relative contribution to the overall score and ascertaining the correct levels of importance can be problematic. Kleist [13] also cautioned that the components in the multiple outcomes should be carefully selected and based on biological plausibility and if surrogate outcomes are used, each one of them should have been properly validated that it correlates with a hard clinical outcome. Ideally, expected frequency of each component outcome should be somewhat similar with extra care in combining hard and soft outcomes, for example, fatal and non-fatal events. He concluded that despite inherent difficulties in using composite outcomes, their potentials are plentiful. They include potentials for requiring a smaller sample size or a shorter completion time of the trial. Physicians also find the composite outcomes provide greater support in their clinical decision making process.

Much research to date concerns how best to assign the weight to each score according to its importance, how best to combine the individual scores into a meaningful overall score and how to modify desirability functions for new situations or applications. In particular, Chen *et al.* [3] proposed the augmented desirability function using the weighted geometric mean for minimizing prediction variances when there are multiple responses.

Desirability functions seem ideally suited for medical studies and analysis of clinical data. Many chronic diseases are increasingly assessed using multiple outcomes to monitor disease progression. The rationale is that multiple ways of ascertaining progression using different outcome measures is likely to result in a more effective assessment of disease management. The multiple outcomes can be discrete or continuous or a mixture of them. If the measure is binary, the response may be improved or not improved. In clinical trials, continuous outcome measures are frequently used either because they arise naturally or because they provide more power than binary outcome measures [2]. There are several such examples of composite measures that incorporate multiple outcomes into one overall measure for evaluating disease progression. For example, composite scores or pooled indices currently used in Rheumatoid Arthritis research were listed in [1]. Fransen and van Riel [8] went further and listed the pooled indices that combine continuous outcome measures developed for rheumatoid arthritis from 1956 to 2007. Some of these composite scores are continuous and they include the disease activity score (DAS), the clinical disease activity index, and the simplified disease activity index. It appears that in rheumatoid arthritis research, the DAS for 28 and 44 joints (DAS28 and DAS44) are among the most popular composite continuous scores for evaluating disease progression [12,20,23]. An example of a binary composite score is the American College of Rheumatology Criteria (ACR) for improvement in rheumatoid arthritis [6]. It has 7 components, requiring improvement in swollen and tender joint counts and at least 20% improvement in 3 of the following components: patient assessment, physician assessment, erythrocyte sedimentation rate, pain scale and functional questionnaire (HAQ). This is the ACR20 improvement criterion.

In practice, the number of components in a composite score varies, typically from 3 to 8. Each component provides a score suggestive of improvement or not and the composite score integrates the subscores into an overall score for disease progression. This overall score can be binary, whether patient has improved or not as in the ACR20 improvement criterion, or a single

number with good clinical interpretation on how much patient has improved, as in the DAS44 composite score.

There is only a handful of papers on use of desirability functions for biomedical studies. Shih *et al*. [22] used desirability functions to titrate and evaluate multi-drug regimens within subjects using a modified evolutionary operation approach and climbed through the dose space to locate dose that patient had an improved response. They used a logistic cumulative distribution function for specifying the desirability function and then combined desirability functions using the geometric mean. Wong *et al*. [24] used the median of measures in selected groups of patients to construct the desirability function for monitoring scleroderma patients' disease progression when outcomes are all continuous. Design problems were also addressed using the desirability function; see [17], where they provided yet another innovative application of the desirability function to design dose response studies for nonlinear models. Additionally, Fransen *et al*. [7] used the medians of 44 expert-rheumatologists' ratings of patients with psoriatic arthritis (PsA) to develop a desirability function for each component measure and combine the desirability scores using the L_p norm. Recently, Helliwell *et al*. [12] developed a composite measure for PsA using the Ankylosing Spondylitis DAS and the desirability function suggested by Fransen *et al*. [7]. After establishing the desirability function cutoffs for disease activity using an online survey technique, these individual functions were then combined into a single measure using the arithmetic mean.

Some key issues in applying desirability functions for medical studies is the availability of a 'gold' standard, the accuracy of each of the constructed desirability functions and the overall desirability function. The shape of the desirability function provides information on how the outcome or its change is interpreted or valued relative to the chosen gold standard, which for clinical applications, is typically some global assessment standard established by an expert pool of physicians. In the above cited work, the desirability function was constructed in an empirical way to reflect consensus from a group of experts and assumed data were available at two time points. Further, when outcomes were perceived to have differential effects on the overall disease progression, the weighted desirability function had weights a priori selected. This suggests that the desirability function constructed in such an empirical way can be problematic in a number of ways: (i) the physicians or experts may not come to a general agreement on what amount of change in the outcome corresponding to what degree of disease progression; (ii) the pool of physicians selected to establish the gold standard and interpretation may be subject to selection bias; (iii) there is no systematic way to select the weight for each outcome even though their perceived importance and usefulness for measuring disease progression may be deduced from literature review, see, for example, [18].

In what is to follow, we propose a data-driven approach to estimate all desirability functions for the various outcomes based on a given gold standard. Specifically, we provide a method to estimate the shape of each desirability function and an appropriate set of weights for the overall desirability function. The advantages of a data-driven approach are that our estimated shapes and weights in the desirability function incorporate cohort characteristics and should reflect disease progression more accurately. In addition, we propose a new approach to using desirability functions with longitudinal data for a more insightful picture of the patient's progress over time relative to the cohort.

The next section describes clinical outcome measures used for assessment of Scleroderma progression in a two-arm randomized trial conducted by Postlewaite *et al.* [19]. Section 3 first gives an overview on the construction of the desirability function and its properties. We then demonstrate our proposed data-driven approach for estimating desirability function shapes and weights and extend the methodology in a novel way to monitor patients' disease progression when there are multiple time points. In Section 4, we apply our proposed method to the real data of Scleroderma. Section 5 offers a discussion.

2. Scleroderma data

Scleroderma is an inflammatory rheumatic disease with multiple disease outcomes. Scleroderma usually affects a large area of skin, lungs, and other internal organs of the patient. We have data from a multicenter, randomized, double-blind, placebo-controlled trial for 168 Scleroderma patients treated with oral type I collagen or placebo. The treatments were administered over a 12-month period with a follow up visit at month 15. Details and results of the trial can be found in [19]. Following Wong *et al*. [24] and Postlethwaite *et al*. [19], five continuous outcome measures are used to evaluate diffuse scleroderma disease progression, including the modified Rodnan total skin score (SKINTOT), the disability index of the health assessment questionnaire (HAQ), the predicted values of the forced vital capacity (FVCP), the patient's global assessments of health (PGA), and the physician overall assessment of disease activity (POA).

Typically, SKINTOT and HAQ are considered to be the most important and serve as two primary outcomes in a Scleroderma clinical trial [19]. SKINTOT is quantitative by assessing 17 body areas with each area receiving a 0–3 score with 0 = normal; 1 = mild thickness; 2 = moderate thickness; 3 = severe thickness, and summing over all the areas. HAQ is a self-reported patient-oriented measure and has four scoring conventions for the disability index questions on a scale of 0–3 with 0 = without any difficulty, 1 = with some difficulty, 2 = with much difficulty, and 3 = unable to do. FVCP is calculated from patient's characteristics and is more fully described in [24]. It takes on positive values with no upper bound. The last two outcome measures are PGA and POA and both aim to track disease progression as judged by the patient alone or by the physician alone. Either one of these may be treated as a gold standard for assessing disease progression in Scleroderma patients in a clinical trial. Both are rated using Visual Analogue Scales on 0–100 mm varying from no activity to extremely high activity.

In this data, we have SKINTOT at months 0, 4, 8, 12, and 15. The maximum score for this measure is 51 with lower values (relative to baseline) indicative of disease progression. We have HAQ data at months 0, 4, 8, 12, and 15 and a lower value of HAQ is suggestive of functional progression. A higher value of FVCP is suggestive that lung functions in the patient are performing better. Such data is expensive to collect and we only have data at months 0 and 12. Both PGA and POA are available at months 0, 4, 8, 12, and 15 in the trial and a lower value is better.

Following Postlethwaite *et al*. [19], the change score for each outcome from baseline to month 12 is used to measure disease activity, calculated by subtracting the score at baseline from the score at the12-month visit. This choice of using the raw difference is the simplest way, and other options of defining change in outcomes are possible. In this trial, Postlethwaite *et al*. [19] showed that oral type I collagen treatment was no more effective than placebo. Accordingly, we combined placebo and treatment groups for measuring patients' disease progression in this study.

3. Methods

3.1 *Desirability function*

The desirability function was first used in the manufacturing sector to combine multiple continuous outcomes into one numeric score as an overall assessment of the quality of the product [10]. Assuming all responses are independent, this approach first creates an individual desirability function d_i to map the ith outcome value y_i to a scalar between 0 and 1, where a higher value of d_i represents greater desirability. This transformation has to be meaningful and make sense in the context of the problem. For instance, if large values of outcome are more desirable, the transformation d_i should be a monotonic increasing function of y_i.

Harrington [10] used exponential functions to transform y_i to d_i, specifically $d_i = \exp(-exp(-y_i))$ for a one-side transformation and $d_i = \exp(-|y_i|^r)$ for a two-sided transformation. Here r is a user-selected shape parameter. Derringer and Suich [5] modified

Harrington's transformations and classified them into three forms: the smaller-the-better (STB), the larger-the-better (LTB), and the nominal-the-best (NTB) types. For the STB type, it is desirable to have the outcome value as small as possible. There is a lower bound L_i and an upper bound U_i for the ith outcome y_i such that it is unacceptable when $y_i > U_i$ and described as perfect or most desirable when $y_i < L_i$. Such a desirability function is defined by

$$d_i = \begin{cases} 1 & \text{for } y_i < L_i, \\ \left(\frac{U_i - y_i}{U_i - L_i}\right)^{r_i} & \text{for } L_i \le y_i \le U_i, \\ 0 & \text{for } y_i > U_i. \end{cases}$$

where $r_i > 0$ is the shape parameter and, like L_i and U_i, is user-selected. The choices of r_i, L_i, and U_i are subjective and their values control the shape of the desirability function. The desirability function is linear when $r_i = 1$. Larger values ($r_i > 1$) signify the importance of being close to L_i. On the other hand, small values ($0 < r_i < 1$) suggest that the outcome does not have to be very close to L_i. When it is desirable for the STB type to have the value of the outcome considerably below U_i, L_i should be closer to U_i; otherwise the range between L_i and U_i can be larger if it is not critical to have the value of the outcome considerably below U_i. The analogous formulas are available for the LTB and the NTB types in [5].

The individual desirability scores $d_i's$ from the various components are then combined into one overall desirability score using the geometric mean $D = (d_1 d_2 \cdots d_m)^{1/m}$, where m is the number of outcomes [10]. The value of D is between 0 and 1. The higher the value of D, the more desirable is the overall product. Clearly, high values of the $d_i's$ result in a high value of D. If all the outcomes are not equally important, a weighted desirability function may be used to reflect the varying contribution from each outcome to the overall desirability score as follows:

$$D_w = (d_1^{w_1} d_2^{w_2} \cdots \mathrm{d}_m^{w_m})^{1/\sum w_i}. \tag{1}$$

The weights satisfy $w_i > 0$ with more important outcomes having larger weights [4]. The D or D_w function has the property that if one characteristic or response is totally unacceptable ($d_i = 0$), then the overall product or process receives a desirability score $D = 0$. This assumes that the particular component measure is extremely important because it can solely decide on the overall score. Such a property may or may not be desirable but whether or not to assign a component score equal to zero has to be carefully weighed in. Alternatively, one can define the overall desirability D as the arithmetic mean or weighted averages of the $d_i's$ with or without some penalty function; see [14–16]. Fransen *et al.* [7] recommended the L_p norm to combine desirability scores in a flexible way after a careful choice of the value for p. This norm is defined by $L_p = ((d_1^p + d_2^p + \cdots + d_m^p)/m)^{1/p}$ and it can be shown that L_0, L_1, L_∞, and $L_{-\infty}$ correspond, respectively, to the geometric mean, arithmetic mean, maximum, and minimum. For example, we note that when $p = 0$,

$$L_p = \left(\frac{1}{m}\sum_{i=1}^{m} d_i^p\right)^{1/p} = \exp\left(\ln\left[\left(\frac{1}{m}\sum_{i=1}^{m} d_i^p\right)^{1/p}\right]\right) = \exp\left(\frac{\ln((1/m)\sum_{i=1}^{m} d_i^p)}{p}\right)$$

and applying the L'Hôpital's rule, one obtains

$$\lim_{p\to 0}\frac{\ln((1/m)\sum_{i=1}^{m} d_i^p)}{p} = \lim_{p\to 0}\frac{((1/m)\sum_{i=1}^{m} d_i^p \ln\ d_i)/((1/m)\sum_{i=1}^{m} d_i^p)}{1}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \ln\ d_i = \ln\left(\prod_{i=1}^{m} d_i\right)^{1/m}$$

and consequently,

$$\lim_{p\to 0} L_p = \exp\left(\ln(\prod_{i=1}^{m} d_i)^{1/m}\right) = \left(\prod_{i=1}^{m} d_i\right)^{1/m}.$$

3.2 *Proposed methods*

3.2.1 *Data-driven approach*

Estimate of the desirability function shape. This parameter is important because different values of this shape parameter signify the degrees of importance of the various amounts of change. For example, a relatively flat desirability function suggests that a relatively large amount of change in the outcome will be required to impress physicians that there is real progression in the patient based on the particular outcome. Physicians however may have their own preferences for the value of the shape parameter and there may not be a consensus on an appropriate value for the shape parameter to use in the study. One way to tackle this issue is to use a data-driven approach to estimate the desirability function shape.

For illustrative purposes, consider that the ith variable has the STB type of desirability function as an example. Given the lower and upper bounds L_i and U_i, the desirability function d_i is a function of the outcome y_i. When $L_i < y_i < U_i$, the desirability function is $d_i = ((U_i - y_i)/(U_i - L_i))^{r_i}$. Suppose that d_g is the given gold standard and rescaled to vary from 0 to 1. In the data-driven approach, we want d_i to approximate d_g as closely as possible. A simple model is

$$d_g = \left(\frac{U_i - y_i}{U_i - L_i}\right)^{r_i} + \varepsilon_i, \tag{2}$$

where ε_i is the approximation error. This is a nonlinear regression problem and we use nonlinear least squares method to estimate the shape parameter r_i. The same procedure can be directly applied to estimate desirability function shapes for the LTB and NTB types when a target value is available. Once the desirability function shape r_i for each outcome is obtained from the data-driven approach and the given gold standard, the desirability score d_i for each outcome can be determined.

Estimate of the desirability function weights. The data-driven approach can be employed to find suitable weights for various outcomes as follows. Given individual desirability scores, $d_1, \ldots, d_m$, from various components, in order to approximate the rescaled gold standard d_g with D_w in (1), we consider the following model

$$d_g = (d_1^{w_1} d_2^{w_2} \cdots d_m^{w_m})^{1/\sum w_i} + \varepsilon. \tag{3}$$

We can fix one of the weights, say $w_1 = 1$, in Equation (3) and use nonlinear least squares method to estimate the other weights w_i. The choice for which one of the weights to drop is arbitrary as it does not affect the results. We can always rescale the weights to make $\sum_{i=1}^{m} w_i = 1$ after the estimation. The estimated weights allow us to compute the value of the overall desirability score D_w in Equation (1). Our experience is that the estimated weights from biomedical data are usually positive. If necessary, we can use constrained nonlinear least squares to ensure that all estimated weights are non-negative.

3.2.2 *Analyzing longitudinal data*

For each time point, let d_{it} be the desirability score for change in each outcome y_i at time t. We define an overall desirability score at time t by $D_t = (d_{1t}^{w_1} \cdots d_{mt}^{w_m})^{1/\sum w_i}$, where w_i is the

estimated data-driven weight of the outcome y_i and $w_i > 0$. For multiple time points, we propose to use the modified L_p norm to combine desirability scores at different time points into one single score. Specifically, suppose we have repeated measurements at k time points $t_1, t_2, \ldots, t_k$ and $D_{t_1}, D_{t_2}, \ldots, D_{t_k}$ are overall desirability scores constructed from each single time point. These are then combined into one composite overall desirability score:

$$D^* = \left[\frac{\lambda_1 \times D_{t_1}^p + \cdots + \lambda_k \times D_{t_k}^p}{\lambda_1 + \cdots + \lambda_k} \right]^{1/p},$$

where λ_j is the weight of time point t_j and $j = 1, \ldots, k$.

4. Application

We apply our proposed methods to the real data of Scleroderma for monitoring disease progression of patients over time. For illustrative purposes, we first monitor patients' disease progression with data at baseline and at month 12 for each outcome. The reason for considering two time points is that SKINTOT and HAQ are available at months 0, 4, 8, 12, and 15 in the trial, but FVCP is only available at months 0 and 12. The desired type and measures of change in each outcome are given in Table 1. In this study, a negative change implies disease progression in SKINTOT (y_1), HAQ (y_2), PGA (y_4), and POA (y_5) while a positive change implies progression in FVCP (y_3).

Following current practice in Scleroderma trials, we select change in one of the outcomes PGA and POA as our gold standard as a reference for mapping changes in other outcomes SKINTOT, HAQ, and FVCP into desirability scores. The bounds L_i and U_i for change in each outcome are chosen based on their observed ranges from baseline to month 12. The first step is to rescale changes in PGA and POA to values between 0 and 1 using the desirability function. In this study, we do not have these desirability functions from experts. For illustrative purposes, we assume $r_4 = r_5 = 1$ in the following analyses. For each patient, we obtain the values of d_4 and d_5 for changes in PGA and POA from baseline to one year. Based on the rescaled scores d_4 and d_5, we estimate the desirability function shapes for changes in SKINTOT, HAQ, and FVCP using the data-driven approach. Using PGA as the gold standard, we obtain $\hat{r}_1 = 0.68$ with the 95% confidence interval (0.6, 0.76) listed in Table 1. Similar calculations are applied to each outcome when POA is chosen as the gold standard. The robustness to extreme outcomes was evaluated and no major changes were observed.

Table 1 also shows that the upper limits of the 95% confidence intervals for estimated desirability function shapes for changes in SKINTOT, HAQ, and FVCP using the gold standard PGA are smaller than 1, but their lower confidence limits for changes in HAQ and FVCP using the

Table 1. Desired type and estimate of the desirability function shape and weight for change in each outcome from baseline to one year using $r_4 = r_5 = 1$.

Calibrating variable (desired type : range)	Change in outcome (desired type: range)	Estimate of shape	95% Confidence interval	Estimate of weight
PGA (STB: [− 67, 81])	SKINTOT (STB: [− 23, 18])	0.68	(0.6, 0.76)	0.45
	HAQ (STB: [− 1.37, 1.63])	0.74	(0.64, 0.84)	0.42
	FVCP (LTB: [− 54, 41])	0.84	(0.74, 0.95)	0.13
POA (STB: [− 76.5, 66])	SKINTOT	1.1	(0.95, 1.26)	0.13
	HAQ	1.15	(1.04, 1.28)	0.68
	FVCP	1.29	(1.15, 1.44)	0.18

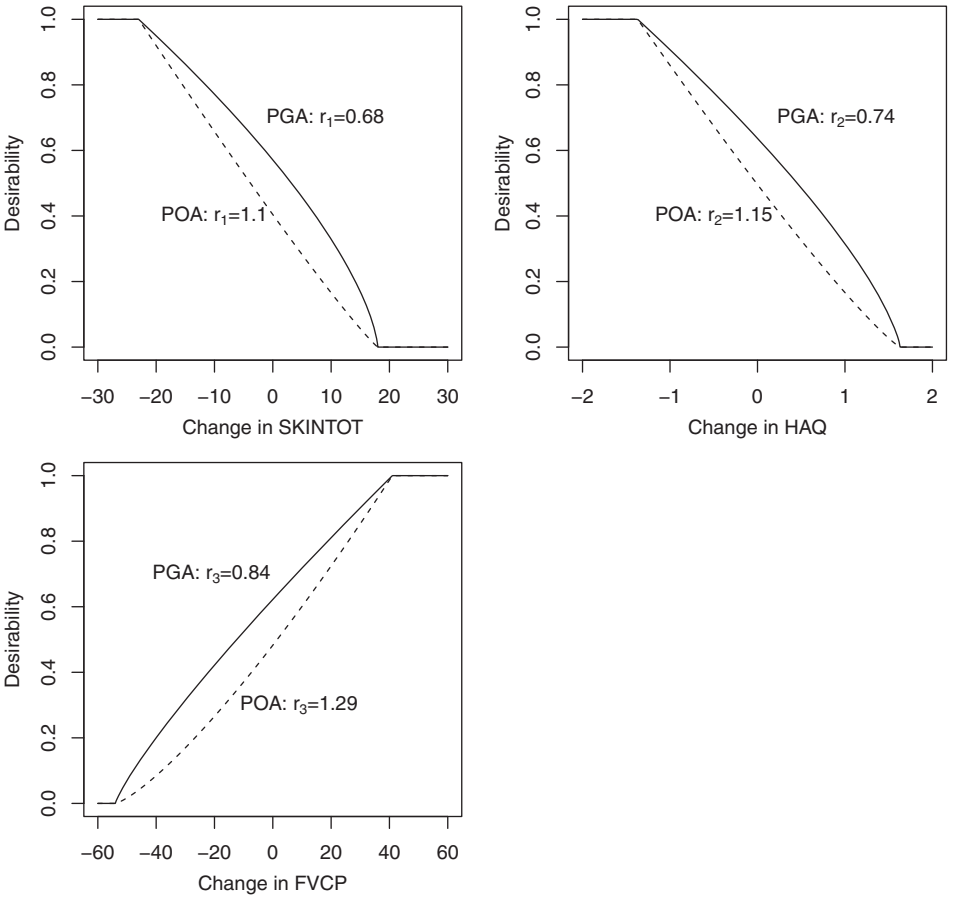

Figure 1. Estimated data-driven desirability function for change in SKINTOT, HAQ, and FVCP from baseline to one year with $r_4 = r_5 = 1$.

gold standard POA are larger than 1. This implies that the gold standard PGA tends to provide smaller estimates of desirability function shapes. We next estimate weights for SKINTOT, HAQ, and FVCP using the data-driven approach. The estimated weights for SKINTOT, HAQ, and FVCP, shown in Table 1, are respectively 0.45, 0.42, and 0.13 when PGA is used as the gold standard, and 0.13, 0.68, and 0.18 when POA is used. Under PGA scale, change in SKINTOT receives the highest weight and is deemed the most important. When POA is used, the corresponding change in HAQ has the largest weight suggesting that HAQ contributes the most to the overall score. Clearly, patients perceive progression in SKINTOT as the major contributing factor to overall disease progression among the various measures considered while the physicians place greatest emphasis on HAQ as the major contributing factor to disease progression. This suggests that PGA and POA may be intrinsically different constructs when they are used to interpret disease progression. The finding here gives support that physicians and patients may have different perception of disease progression.

Figure 1 shows that the estimated data-driven desirability functions for change in SKINTOT, HAQ, and FVCP from baseline to one year with $r_4 = r_5 = 1$. We observe that PGA and POA have different appreciations of an increase in change in the three outcomes. Specifically, if PGA is used as the gold standard, we have higher desirability scores for SKINTOT, HAQ, and FVCP. This means that using PGA as the gold standard, a small change or no change in the three

Table 2. Estimate of the desirability function shape and weight for change in SKINTOT and HAQ from baseline to each time point.

Time point	Calibrating variable	Change in outcome	Range	Estimate of shape	95% Confidence interval	Estimate of weight
Month 4	PGA	SKINTOT	[− 14, 19]	1.47	(1.31, 1.64)	0.54
		HAQ	[− 1.62, 2.25]	1.28	(1.14, 1.43)	0.46
	POA	SKINTOT	[− 14, 19]	1.47	(1.3, 1.65)	0.27
		HAQ	[− 1.62, 2.25]	1.28	(1.17, 1.39)	0.73
Month 8	PGA	SKINTOT	[− 21, 13]	0.81	(0.72, 0.91)	0.68
		HAQ	[− 1.62, 1.5]	0.91	(0.79, 1.05)	0.32
	POA	SKINTOT	[− 21, 13]	1.02	(0.88, 1.17)	0.22
		HAQ	[− 1.62, 1.5]	1.2	(1.08, 1.33)	0.78
Month 12	PGA	SKINTOT	[− 23, 18]	0.68	(0.6, 0.76)	0.58
		HAQ	[− 1.37, 1.63]	0.74	(0.64, 0.84)	0.42
	POA	SKINTOT	[− 23, 18]	1.1	(0.95, 1.26)	0.26
		HAQ	[− 1.37, 1.63]	1.15	(1.04, 1.28)	0.74
Month 15	PGA	SKINTOT	[− 25, 21]	0.81	(0.72, 0.91)	0.67
		HAQ	[− 1.63, 1.5]	0.65	(0.56, 0.75)	0.33
	POA	SKINTOT	[− 25, 21]	1.08	(0.94, 1.24)	0.25
		HAQ	[− 1.63, 1.5]	0.91	(0.81, 1.03)	0.75

outcomes is deemed as more indicative of some progression than if POA is used as the gold standard. More specifically, Figure 1 suggests that no change in any one of the three outcomes would each receive a desirability score of about 0.4 to 0.5 when POA is used as the gold standard versus a desirability score of about 0.6 if PGA is used as the gold standard. These scores translate to a rating of 'acceptable but poor' interpretation according to Harrington's guidelines [10]. The upshot is that while the values for the desirability scores are different from different gold standards, the qualitative interpretation of disease progression is similar.

Next, for illustrative purposes, we drop FVCP and work with only SKINTOT and HAQ from multiple visits. Table 2 shows their estimated desirability function shapes and corresponding 95% confidence intervals, as well as their estimated weights over time. At month 4, the estimates of shape are all larger than 1 and are similar using either PGA or POA as the gold standard. At months 8, 12, and 15, the estimates of shape under PGA scale are smaller than the estimates under POA scale. This implies that patients and physicians had similar views in disease progression at month 4 but their views diverged over time. The patients appeared to be more positive than the physicians at months 12 and 15. In addition, we observe that larger weights are obtained for SKINTOT under PGA and smaller weights are obtained for SKINTOT under POA at each time point.

Table 3 shows the changes and estimated data-driven desirability scores in SKINTOT and HAQ from baseline to each time point for the patient with id = 16. With $p = 1$ and equal weights, the patient has a composite overall score of $D^* = (0.47 + 0.51 + 0.65 + 0.85)/4 = 0.62$ under PGA. In practice, current drug treatments for Scleroderma are slow acting and not expected to have an effect on the patient until about four or more months later. This means that change measurements from patients at four months or earlier should receive small weights. Measurements after six months should receive larger weights increasingly for change measurements collected over a longer period of time. For this reason, it seems more appropriate to assign increasing weights such as $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 2, 4, 8)$ for this study. With this choice of weights and $p = 1$, the composite overall score becomes $D^* = (1 \times 0.47 + 2 \times 0.51 + 4 \times 0.65 + 8 \times 0.85)/(1 + 2 + 4 + 8) = 0.726$ under PGA. This suggests that the choice of other increasing weights over time does not markedly affect the results. Using this patient, we further illustrate two important features of the composite overall

Table 3. Desirability score of change in SKINTOT and HAQ from baseline to each time point for the patient (id = 16).

Calibrating variable	Time point	Change in SKINTOT	Desirability of SKINTOT	Change in HAQ	Desirability of HAQ	D_w
PGA	Month 4	0	0.45	0	0.5	0.47
	Month 8	− 3	0.54	0.25	0.44	0.51
	Month 12	− 12	0.81	0.5	0.49	0.65
	Month 15	− 13	0.78	− 1.63	1	0.85
POA	Month 4	0	0.44	0	0.5	0.48
	Month 8	− 3	0.46	0.25	0.34	0.36
	Month 12	− 12	0.71	0.5	0.32	0.4
	Month 15	− 13	0.72	− 1.63	1	0.92

Table 4. Measures of the desirability scores at each time point and the composite overall desirability D^* using $p = 1$ for collagen and placebo groups with the set of weights $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 2, 4, 8)$ for months 4, 8, 12, and 15.

		Collagen group		Placebo group			
Calibrating variable	Time point	Mean of desirability	sd of desirability	Mean of desirability	sd of desirability	t statistics of desirability scores from 2 groups	p-Value
PGA	Month 4	0.47	0.13	0.49	0.11	− 0.99	0.32
	Month 8	0.49	0.12	0.49	0.15	0.05	0.96
	Month 12	0.61	0.13	0.6	0.16	0.43	0.67
	Month 15	0.61	0.12	0.57	0.15	1.24	0.22
	Overall	0.59	0.1	0.57	0.14	1	0.32
POA	Month 4	0.46	0.13	0.49	0.11	− 1.45	0.15
	Month 8	0.4	0.13	0.41	0.14	− 0.57	0.57
	Month 12	0.46	0.15	0.47	0.16	− 0.35	0.72
	Month 15	0.5	0.14	0.48	0.16	0.53	0.6
	Overall	0.49	0.12	0.47	0.14	0.54	0.59

desirability scores. First, we illustrate that the score of D^* does not change greatly after having a low score at an early time point. Suppose month 4 had $D_{t1} = 0.1$ (instead of 0.47) under PGA, the score of D^* would be 0.70. That is, a score near 0 at the beginning will not affect the final result too much. Second, we illustrate how to deal with missing values without imputation in practice. Suppose that data at month 8 are missing under PGA and we only have the scores at months 4, 12, and 15 for this patient. We can simply set the weight $\lambda_2 = 0$ for month 8 and keep the weights for other time points. Then the composite overall score can be computed as $D^* = (1 \times 0.47 + 4 \times 0.65 + 8 \times 0.85)/(1 + 4 + 8) = 0.76$.

The desirability scores can be used in analyzing longitudinal data. For illustration, we consider using the desirability scores to test whether the collagen and placebo groups were significantly different. Table 4 shows the means and standard deviations of the desirability scores for two groups, t statistics and p-values at each time point, as well as the measures of the composite overall desirability D^* using $p = 1$ and weights $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 2, 4, 8)$. The large p-values indicate no difference between collagen and placebo groups at each time point or across all time points. This reaffirms the conclusions from [19] that the collagen group did not fare significant better over the placebo group over time.

5. Discussion

Measuring patients' disease progression with multiple outcomes is increasingly common, but challenges arise when we try to combine all these outcomes into a single meaningful progression

score that incorporates multiple outcome measures appropriately. We propose the data-driven approach to construct desirability functions that can potentially highlight selection bias and subjective judgments. Specifically, unlike previous approaches, we estimate the shapes of the desirability functions and also the weights for all outcomes based on the given gold standard and the observed data. The upshot is that each patient has a meaningful overall progression score that facilitates clinical interpretation. In addition, we show how additional information from a longitudinal study can be integrated into the single desirability score in a novel way for a more insightful picture of patients' progress over time.

A salient finding from our analyses is that our estimates of shapes and weights of the desirability functions depend on the choice of the gold standard. When PGA is used as the gold standard, we always have a larger desirability score for the yearly change in SKINTOT, HAQ, and FVCP than if we use POA as the gold standard. Even if the values for the desirability scores are different from different gold standards, their qualitative interpretation of disease progression is similar in this study. Overall, we find that scleroderma patients in this trial do not show a significant progression over time.

The use of the geometric mean to combine the various desirability scores means that the overall desirability score is 0 as soon as one of the sub scores is 0. An alternative way may use the arithmetic mean, where poor scores on one component can be compensated by good scores on another component. Our proposed method with the modified L_p norm for analyzing longitudinal data works when a patient has a desirability score near 0 at the beginning of the trial but then markedly improves. Poor scores at early time points can be compensated by good scores at later time points in D^*. However, for the multiplicative approach, this compensation is only possible on relatively high scores at most time points. As scores at some time points become lower, it is more difficult to compensate at all and usually result in a poor composite score.

Missing data are common in clinical trials. In this study, we work with patients with complete data only. We further conduct multiple imputations using package (mice) in *R* to generate 10 complete data sets. The p values for testing collagen and placebo groups are all above 0.30 using the composite overall desirability scores, which is consistent with Table 4. In practice, there is a need for a doctor to deal with missing data for an individual patient without multiple imputation. With our proposed method for analyzing longitudinal data, a doctor can simply use the available data by setting $\lambda_j = 0$ if the outcomes at time t_j are missing in D^*.

In longitudinal analysis, weights may be chosen proportional to the distance between baseline and at user-selected time point, with larger weights for longer distances. This is particularly appropriate for rheumatic diseases where the treatment is usually slow-acting, meaning that effects of the drug will not manifest themselves until a few months later or after the accumulated doses in the patient exceeds some threshold. However, we do not incorporate the correlation in the desirability scores across time in our proposed method for analyzing longitudinal data.

In this study, we do not have information to derive the desirability functions and weights from a panel of experts. Our estimated data-driven shapes and weights of the desirability functions depend on rescaled PGA and POA, which are based on $r_4 = r_5 = 1$. The values of r_4 and r_5 are chosen in an empirical way in this study. Accordingly, we compare the results under PGA and POA scales.

In summary, our work appears to be the first to discuss desirability function and its potential use in longitudinal studies. Our work provides two innovations for the use of desirability functions. First, we used a data-driven approach to estimate the desirability function and second, we are the first to apply in a novel way how desirability functions can be used in the analysis of longitudinal data. The method is relatively simple and so may appeal to researchers who want more insight how different component measures affect the overall disease progression of the patients. Our method also incorporate missing data in a simple way but more sophisticated methods can be used. For example, one may wish to incorporate correlations among the multiple outcomes in

ways similar to recent ideas proposed in the literature [25,21]. Our application was for rheumatic diseases but the method is general and can be used to broadly analyze other longitudinal studies as well. Of course, we do not suggest that this approach supplants conventional statistical methods for longitudinal analyses but use the desirability approach as an alternative and simple method to possibly gain insights beyond those obtained from standard statistical analyses.

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research of Wong reported in this publication was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health [grant number R01GM107639].

References

[1] M. Boers and P. Tugwell, *The validity of pooled outcome measure (indices) in rheumatoid arthritis clinical trials*, J. Rheumatol. 20 (1993), pp. 568–574.
[2] G.F. Borm, G.J. van der Wilt, J.A.M. Kremer, and G.A. Zielhuis, *A generalized concept of power helped to choose optimal endpoints in clinical trials*, J. Clin. Epidemiol. 60 (2007), pp. 375–381.
[3] H.W. Chen, W.K. Wong, and H. Xu, *An augmented approach to the desirability function*, J. Appl. Stat. 39 (2012), pp. 599–613.
[4] G.C. Derringer, *Balancing act: Optimizing a product's properties*, Qual. Prog. 27 (1994), pp. 51–58.
[5] G.C. Derringer and R. Suich, *Simultaneous optimization of several response variables*, J. Qual. Technol. 12 (1980), pp. 214–219.
[6] D.T. Felson, J.J. Anderson, M. Boers, C. Bombardier, D. Furst, C. Goldsmith, L.M. Katz, R. LightfootJr.H. Paulus, V. Strand, P. Tugwell, M. Weinblatt, H.J. Williams, F. Wolfe, and S. Kieszak, *American college of rheumatology. Preliminary definition of improvement in rheumatoid arthritis*, Arthritis Rheum. 38 (1995), pp. 727–735.
[7] J. Fransen, A. Kavanaugh, and G.F. Borm, *Desirability scores for assessing multiple outcomes in systemic rheumatic diseases*, Comm. Statist. Theory Methods 38 (2009), pp. 3461–3471.
[8] J. Fransen and P.L. van Riel, *Outcome measures in inflammatory rheumatic diseases*, Arthritis Res. Ther. 11 (2009), pp. 244.
[9] P.L. Goethals and B.R. Cho, *Extending the desirability function to account for variability measures in univariate and multivariate response experiments*, Comput. Eng. 62 (2012), pp. 457–468.
[10] E.C. Harrington, *The desirability function*, Ind. Qual. Control 4 (1965), pp. 494–498.
[11] Z. He, P.F. Zhu, and S.H. Park, *A robust desirability function method for multi-response surface optimization considering model uncertainty*, European J. Oper. Res. 221 (2012), pp. 241–247.
[12] P.S. Helliwell, O. Fitzgerald, and P.J. Mease, *Development of composite measures for psoriatic arthritis: A report from the GRAPPA 2010 annual meeting*, J. Rheumatol. 39 (2012), pp. 398–403.
[13] P. Kleist, *Composite endpoints for clinical trials*, Int. J. Pharm. Med. 21 (2012), pp. 187–198.
[14] J.F. Kros and C.M. Mastrangelo, *Comparing methods for the multi-response design problem*, Qual. Reliab. Eng. Int. 17 (2001), pp. 323–331.
[15] A. Mandal, K. Johnson, C.F.J. Wu, and D. Bornemeier, *Identifying promising compounds in drug discovery: Genetic algorithms and some new statistical techniques*, J. Chem. Inf. Model. 47 (2007), pp. 981–988.
[16] F. Ortiz, J.R. Simpson, J.J. Pignatiello, and A. Heredia-Langner, *A genetic algorithm approach to multiple-response optimization*, J. Qual. Technol. 36 (2004), pp. 432–450.
[17] S.M. Parker and C. Gennings, *Penalized locally optimal experimental designs for nonlinear models*, J. Agric. Biol. Environ. Stat. 13 (2008), pp. 334–354.
[18] H. Paulus, K.J. Bulpitt, B. Ramos, G. Park, and W.K. Wong, *Relative contributions of the components of the American College of Rheumatology 20% criteria for improvement to responder status in patients with early seropositive rheumatoid arthritis*, Arthritis Rheum. 43 (2000), pp. 2743–2750.

[19] A.E. Postlethwaite, W.K. Wong, P. Clements, S. Chatterjee, B.J. Fessler, A.H. Kang, J. Korn, M. Mayes, P.A. Merkel, J.A. Molitor, L. Moreland, N. Rothfield, R.W. Simms, E.A. Smith, R. Spiera, V. Steen, K. Warrington, B. White, F. Wigley, and D.E. Furst, *A multicenter, randomized, double-blind, placebo-controlled trial of oral type I collagen treatment in patients with diffuse cutaneous systemic sclerosis: I. Oral type I collagen does not improve skin in all patients, but may improve skin in late-phase disease*, Arthritis Rheum. 58 (2008), pp. 1810–1822.

[20] M.L. Prevoo, M.A. van't Hof, H.H. Kuper, M.A. van Leeuwen, L.B. van de Putte, and P.L. van Riel, *Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis*, Arthritis Rheum. 38 (1995), pp. 44–48.

[21] A. Salmasnia, R.B. Kazemzadeh, and M.M. Tabrizi, *A novel approach for optimization of correlated multiple-responses based on desirability function and fuzzy logics*, Neurocomputing. 91 (2012), pp. 56–66.

[22] M. Shih, C. Gennings, V.M. Chinchilli, and W.H. Carter, Jr, *Titrating and evaluating multi-drug regimens within subjects*, Stat. Med. 22 (2003), pp. 2257–2279.

[23] D.M. van der Heijde, M.A. van't Hof, P.L. van Riel, L.A. Theunisse, E.W. Lubberts, M.A. van Leeuwen, M.H. van Rijswijk, and L.B. van de Putte, *Judging disease activity in clinical practice in rheumatoid arthritis: First step in the development of a disease activity score*, Ann. Rheum. Dis. 49 (1990), pp. 916–920.

[24] W.K. Wong, D.E. Furst, P.J. Clements, and J.B. Streisand, *Assessing disease progression using a composite endpoint*, Stat. Methods Med. Res. 16 (2007), pp. 31–49.

[25] F.C. Wu, *Optimization of correlated multiple quality characteristics using desirability functions*, Qual. Eng. 17 (2005), pp. 119–126.