VALUE IN HEALTH ∎ (2017) ∎∎∎–∎∎∎

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Choice Defines Value: A Predictive Modeling Competition in Health Preference Research

Michał Jakubczyk, PhD[1], *Benjamin M. Craig, PhD*[2,*], *Mathias Barra, PhD*[3], *Catharina G.M. Groothuis-Oudshoorn, PhD*[4], *John D. Hartman, MA*[5], *Elisabeth Huynh, PhD*[6], *Juan M. Ramos-Goñi, MSc*[7], *Elly A. Stolk, PhD*[7], *Kim Rand, PhD*[3,8]

[1]*Decision Analysis and Support Unit, SGH Warsaw School of Economics, Warsaw, Poland;* [2]*Department of Economics, University of South Florida, Tampa, FL, USA;* [3]*Health Services Research Centre, Akershus University Hospital, Lørenskog, Norway;* [4]*Department of Health Technology and Services Research, University of Twente, Twente, The Netherlands;* [5]*Haas Center for Business Research and Economic Development, University of West Florida, Pensacola, FL, USA;* [6]*Institute for Choice, University of South Australia, Adelaide, Australia;* [7]*EuroQol Research Foundation, Rotterdam, The Netherlands;* [8]*Department of Health Management and Health Economics, University of Oslo, Oslo, Norway*

ABSTRACT

Objective: To identify which specifications and approaches to model selection better predict health preferences, the International Academy of Health Preference Research (IAHPR) hosted a predictive modeling competition including 18 teams from around the world. **Methods:** In April 2016, an exploratory survey was fielded: 4074 US respondents completed 20 out of 1560 paired comparisons by choosing between two health descriptions (e.g., longer life span vs. better health). The exploratory data were distributed to all teams. By July, eight teams had submitted their predictions for 1600 additional pairs and described their analytical approach. After these predictions had been posted online, a confirmatory survey was fielded (4148 additional respondents). **Results:** The victorious team, "Discreetly Charming Econometricians," led by Michał Jakubczyk, achieved the smallest χ^2, 4391.54 (a predefined criterion). Its primary scientific findings were that different models performed better with different pairs, that the value of life span is not constant proportional, and that logit models have poor predictive validity in health valuation. **Conclusions:** The results demonstrated the diversity and potential of new analytical approaches in health preference research and highlighted the importance of predictive validity in health valuation.

Keywords: discrete choice experiments, EQ-5D, health preference research, QALY.

Introduction

Crowdsourcing is the process of obtaining services, ideas, or content by soliciting contributions from a large group of people rather than by relying on a single person or a handful of collaborators. By gathering the ideas of multiple independent teams, such a communal endeavor fosters greater creativity and tends to achieve a wider range of possible solutions and perspectives. The International Academy of Health Preference Research (IAHPR) hosted a predictive modeling competition designed on the premise that the community of health preference researchers is diverse in modeling expertise and perspectives [1]. Instead of relying on convention, peer review, or theoretical assumptions, the competition described in the present article produced a diversity of analytical approaches by striving for greatest predictive validity.

Health preference research (HPR) is a scientific enterprise: specifications are devised, hypothesized, and tested. Its mantra, "choice defines value," refers to the importance of choice evidence to understand the value people place on health and health care [2]. Nevertheless, by convention within HPR, researchers typically conduct just one preference study, estimate just one analytic specification, and promote its implementation without confirmation. It seems misguided to ground health policy decisions on preliminary evidence acquired and presented from the perspective of a single team. More troubling is the approach of researchers who estimate multiple specifications and cherry-pick their results (as in data mining) [3]. In clinical trials, analysis plans must be formally registered before collecting and examining the data [4], and the results are typically confirmed by multiple teams before putting them into practice. For this purpose, the IAHPR launched the Health Preference Study and Technology Registry

Conflicts of interest: The authors have indicated that they have no conflicts of interest with regard to the content of this article.

* *Address correspondence to:* Benjamin M. Craig, Department of Economics, University of South Florida, 4202 E. Fowler Avenue, CMC206A, Tampa, FL 33620.

E-mail: bcraig@usf.edu

1098-3015$36.00 – see front matter Copyright © 2017, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

http://dx.doi.org/10.1016/j.jval.2017.09.016

(hpstr.org). HPR teams can post their analytical plans on this registry before data collection.

In addition to demonstrating a diversity of analytical approaches, this competition was designed to promote scientific rigor in HPR by having multiple teams compete and then judging the winner on the basis of confirmatory, rather than exploratory, results. To our knowledge, this is the first predictive modeling competition in HPR. Improving the understanding of how people make choices in experimental settings is particularly important in HPR, because health is not bought and sold openly. Therefore, to understand the value of various attributes thereof, health preference researchers conduct surveys using such elicitation techniques as paired comparisons [5].

For the predictive modeling competition reported here, data on paired comparisons from an exploratory survey were distributed to multiple teams so that each team might apply its own modeling specifications independently. Using its findings, each team submitted predictions for a second, confirmatory set of paired comparisons. After their predictions had been posted publicly, a confirmatory survey was fielded and the teams' submissions were ranked in accordance with their predictive validity (smallest to largest χ^2). Although the competition has only one winner, this crowdsourcing endeavor was also designed to yield benefits more generally to the HPR community: by promoting greater understanding of the merits underlying alternative modeling specifications, promoting the importance of predictive validity in HPR, and demonstrating the diversity of analytical approaches among HPR researchers.

Methods

Team Registration

In March 2016, Drs. Craig and Rand-Hendriksen distributed an announcement inviting all interested teams to participate in the predictive modeling competition [1]. By April, 18 teams had registered using a brief form on the IAHPR Web site (no exclusions) that asked five questions pertaining to 1) conditional agreement for teams; 2) the names of the team and team leader, and the number of co-investigators; 3) the names of the co-investigators; 4) experience; and 5) invoicing. By May, all registered teams had received the exploratory data and a sample submission. By July, 8 of the 18 teams had submitted their forms and predictions and were paid $2500 (see Appendix in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.016). In September, the victorious team received a small trophy at the 2016 EuroQol Plenary in Berlin and lead authorship of this article in concordance with the Vancouver criteria [6].

Task and Pair Selection

The design of the paired comparisons (see Fig. 1 for an example) was largely based on the recent protocols for the valuation technology developed by the EuroQol Group (EQ-VT) [7]. The wording differed from the EQ-VT in four ways: 1) Because it was designed to elicit preferences, not judgments, the competition survey instrument asked "Which do you prefer?" instead of "Which is better?" 2) The labels "A" and "B" were dropped, because they might imply rank; 3) The differentiating attributes and numbers were bolded; and 4) Each description included the timing and duration of problems (e.g., "Starting today, [x] years with health problems: [health state] then die ([x] years from today)").

The set of 1560 pairs in the exploratory survey was based on the 196 pairs in the EQ-VT. Every pair had two health descriptions, each of which included five problems based on the five-level EuroQol five-dimensional questionnaire (EQ-5D-5L; mobility [MO], self-care [SC], usual activities [UA], pain/discomfort [PD], and anxiety/depression [AD]). Each problem was characterized as being at one of five possible levels (none [level 1], slight, moderate, severe, and unable/extreme [level 5]). As a shorthand notation, the five problems are standardly characterized as a vector of five numbers (e.g., Fig. 1 includes 33333). The problems based on the 196 pairs of the EQ-VT (and 4 ancillary pairs) had durations in four different temporal units (days, weeks, months, and years), creating 800 efficient pairs.

In addition to the 800 efficient pairs, 760 time trade-off (TTO) pairs of identical structure were included. In TTO, pairs are distinguished by having one health description that involves no health problems (i.e., 11111) and a longer life span (e.g., Fig. 1), like a conventional TTO task. To select the TTO pairs, 38 descriptions were selected from the efficient pairs, included durations in four different temporal units, and paired with five durations with no health problems ($38 \times 4 \times 5 = 760$). Forty of the TTO pairs included "immediate death."

The set of 1600 pairs in the confirmatory survey included 800 efficient pairs as well as 800 TTO pairs (including 40 with immediate death). These choice sets were created using a similar process, albeit with some important differences. Unlike the previous set, which was based on the 196 EQ-VT pairs, the process began by selecting health descriptions that commonly occur in clinical data and limiting the candidate set of pairs to just these empirically observed combinations. The motivation for emphasis on prevalent outcomes is that health preference estimates are commonly applied to summarize health outcomes captured in clinical trials as a means to better inform medical recommendations and resource allocation decisions (e.g., cost-utility analyses). After combining the problems in these prevalent descriptions with durations to create a candidate set, a software package (Ngene [ChoiceMetrics, Sydney, Australia]) was used to select a subset of 200 pairs by D-efficiency, which were combined with the four

Fig. 1 – Example of a paired comparison.