Journal of Clinical & Experimental Cardiology
ISSN: 2155-9880

Kerimi et al., J Clin Exp Cardiolog 2017, 8:3
DOI: 10.4172/2155-9880.1000509

Research Article Open Access

Statistical and Psychological Models of Doctors' Judgments of Heart Patients

Neda Kerimi[1]*, Lars Backlund[2], Ylva Skaner[2], Lars-Erik Strender[2] and Henry Montgomery[1]

[1]*Department of Psychology, Stockholm University, Sweden*
[2]*Center for Family and Community Medicine, Karolinska Institute, Sweden*

Abstract

Using participant data from the medical domain, the robustness of logistic regression (LR) with different cue inclusion levels and two fast and frugal (F&F) models in terms of predictive accuracy and frugality were tested. Two data sets based on judgments of verbally described patients were used: Heart failure (66 analysts), and Hyperlipidemia (38 analysts). In both data sets, when the models were cross-validated, there was a significant decrease in predictive accuracy for all models, especially when all cues were used in LR. The other models had about equal predictive accuracy, also when comparisons were made with actual diagnoses, with a slight advantage for LR in the Heart failure study. LR using the 5% inclusion level was more frugal than F&F. These results emphasize the importance of using cross-validation and of choosing the proper significance levels for cue inclusion and when comparing different judgment models in medical decision making and other fields.

Keywords: Logistic regression; Fast and frugal; Cross-validation; Medical decision making

Introduction

It is common practice to use different regression models such as logistic regression (LR) to capture judgment policies and create normative models of human decision making behavior [1,2]. A recent review of how LR models could be applied for capturing medical decision making is given by Hamm and Young [3]. Critics, however, mean that human reasoning is not like regression and that other models such as fast and frugal (F&F) models are a better option for judgment analysis [4,5]. In this study, we will compare these two approaches in different respects. Comparing these two types of models is not new. However, we argue that the comparisons have not been made fairly because of several factors that will be described later in this paper.

Capturing human judgment policies is called judgment analysis. One standard approach, in medical judgment analysis, has been to present medical doctors with made-up patient cases where the doctor's task is to make a clinical decision. The doctors' decisions are then compared with and predicted from a regression model applied to the same cases. The outcome of the regression model is different weighs of the cues (that is, the weights of different pieces of information provided in the cases) and the regression model consists of the cues with the highest weights. According to F&F models, decision makers do not integrate, nor search all the provided information (i.e. the cues). Instead, they base their decision on one cue, hence the name frugal [6-10]. Marewski and Gigerenzer discuss how F&F models could be used in the medical domain. One of the most well known F&F models is the take the best heuristic: cues are searched in order of their importance until one discriminates, then search stops and all other cues are ignored [5]. The take the best heuristic has been operationalized in terms of the matching heuristic [6,11] that will be focused in the present study.

In the matching heuristic, first a critical value for each cue is identified as the cue value that has the highest frequency of positive diagnosis or a decision to use medical treatment. For example, if the doctor prescribed heart medicine to 60 males and 40 females then the critical value of the cue gender is male (coded as 1, whereas female is coded as 0) because there are more prescriptions of heart medicine associated with male patients than it is with female patients. Second, cue validity is identified as the proportion of cases with the critical value that is associated with a positive diagnosis or a decision to use medical treatment (= 60/100 in our example). Third, the cue validities are ordered in accordance with their validity as defined in the preceding step, and this order indicates the order the cues are searched by the model. Finally, the smallest number of cues that lead to the best prediction is chosen as the number of cues in the model. Backlund, Bring, Skånér, Strender, and Montgomery [12] tested an extended version of the matching heuristic where cue validities were not only based on the proportion of correctly predicted positive responses (diagnose or prescriptions), but also on the proportion of correctly predicted negative responses (non-diagnoses or non-prescriptions).

Empirical results are seemingly contradictory of whether F&F models provide higher predictive accuracy than LR models [6,10,12-16]. Two groups of studies can be discerned that may be denoted as ecological and judgment studies, respectively. The former group of studies concern ecological relations such as cities being state capitals or not depending on having a soccer team or not [7,13,15]. Thus, in these studies not only the independent variables, but also the dependent variables denoted ecological or actual facts, rather than human judgments or decisions. The research question has been to find out how well LR and F&F models, respectively, can predict the dependent variable (e, g, whether city A has bigger population than city B) from cues in the ecology (independent variables, e. g. if the cities are state capitals or not, have soccer teams or not). To address the research question, cross validation is typically used to compare the validities of the models. The models are first fitted to data from a subset of the whole sample (fitting set). The fitted models are then used for predictions in remaining sample, the prediction sample.

In all ecological studies known to us, LR has been fitted to all

***Corresponding author:** Neda Kerimi, Professor, Department of Psychology, Stockholms University, Frescati Hagväg 14, Stockholm, 10691, Sweden, Tel: 46709812014; E-mail: hmy@psychology.su.se

Received October 26, 2016; **Accepted** March 20, 2017; **Published** March 25, 2017

Citation: Kerimi N, Backlund L, Skaner Y, Strender LE, Montgomery H (2017) Statistical and Psychological Models of Doctors' Judgments of Heart Patients. J Clin Exp Cardiolog 8: 509. doi:10.4172/2155-9880.1000509

available cues (e.g., ten cues for inferring which of two German cities that has larger population; Gigerenzer & Todd) [13]. Thus, it has not been taken into account whether a cue contributes significantly or not to the relationship between the cues (independent variables) and the dependent variable, which means that some cues in the fitted model will be based on noise in the data. By contrast, F&F models use a very limited number of cues (e.g., three cues in the study by Gigerenzer and Todd [13]) in the model. Not surprisingly, the fit of LR is substantially reduced, but not of F&F, when applied in the prediction set as a result of over-fitting (too many cues used) in the fitting set [17] and is typically found to give lower predictive accuracy in the prediction set than is true for F&F models.

A strength of the ecological group of studies is that they use cross validation, which adds to the validity of the results. However, the validity may be diminished by the fact that when all cues are used, LR may be more sensitive to over-fitting than is true for F&F models [18]. For example, Czerlinski et al. [13] compared F&F and LR models by means of cross-validation, and found that LR was more over-fitted than F&F models [19,20]. This finding was paralleled by a superior predictive accuracy of some F&F-models, particularly the take the best heuristic [5]. Martignon et al. [15] found evidence for over-fitting in the LR models when comparing LR to two F&F models in a cross-validation study.

In addition, it is problematic to generalize the results from these studies to situations where the dependent variable corresponds to judgments (e.g., human forecasts of sports results or of economic outcomes) rather than ecological facts (e.g., actual sports results or economic outcomes). Thus, the extent to which the fitted models are applicable in human judgment policies is unknown.

The other group of studies – the judgment studies – is the mirror image of the ecological studies; instead, the dependent variable is judgments data rather than ecological facts [6,14,16,19]. As far as we know, no cross-validation has been used in this group of studies for comparing the validity of LR and F&F models. The closest data we have found is Dhami's [20,21] research, where the matching heuristic, in a cross-validation study, was compared to Franklin's rule (a linear compensatory rule involving cue weights that are pre-determined rather than fitted by means of regression analysis). Thus, the models are mostly compared in terms of their fit to a given sample of data rather than cross-validated to a new set of data. Additionally, LR has been based only on cues that contribute significantly to the predicted variance, with p being .05 or .10). The typical result is that LR yields slightly higher or equal predictive accuracy as compared to F&F models. On the other hand, F&F models have been found to be more frugal by using fewer cues in the model than is the case for the LR models. (Note that F&F models could involve checks of more than one cue although only one cue is critical for the decision).

Present study

The present study combines the strengths of the ecological and judgment studies, respectively, in an attempt to make a fairer test of the validity of LR and F&F models than has been done in earlier studies. The data in present study come from the medical domain and include both judgments from individual doctors on a number of patient cases as well as actual diagnoses of the same patients. This makes it possible to model the doctors' judgments as well as the ecological relation to actual diagnoses with the cue profiles in patient cases as independent variables in both cases. In all, five models were tested: Three LR models each corresponding to a certain level of cue inclusion ($p<0.05$, denoted as LR 5, $p<0.10$, denoted as LR 10, and inclusion of all cues, denoted as LR Enter, respectively) and two F&F heuristics, the matching heuristic reported by Dhami and Harries [6], denoted as F&F 1, and extended matching heuristic reported by Backlund et al. [12], denoted as F&F 2). Moreover, we used cross-validation for testing the validity of fitted models by first fitting the models to a subset of the patient cases (fitting set) and then testing the predictive accuracy of the fitted models for the remaining patient cases (prediction set). In order to control for and assess the importance of over-fitting in LR, the predictive accuracy of three different more or less frugal LR-models that had been fitted in the fitting set was compared in the prediction set. Cross-validation was also used for examining the frugality of the models (i.e., the number of cues used in the models). Frugality was not only assessed in terms of number of cues in the models, but also in terms of how many cues in the models that were actually utilized when the search process had come to an end. For example, only one cue is utilized if the cue with the highest cue validity (i.e., the first cue that is checked) discriminates between prescription and non-prescription of a drug.

Method

Data sets

The fitting and testing of models on doctors' diagnose, actual diagnoses, and prescription decisions were based on the same data that Backlund et al. [12]) used in their studies as well as on some additional data. Backlund et al. used two data sets, one on Heart failure diagnosis (27 general practitioners) and one on Hyperlipidemia treatment (38 general practitioners). The new data set, concerned Heart failure diagnosis and were identical to the data set from Backlund et al. in data collection procedure. There were no significant differences in terms of fit and predictive accuracy between the heart failure data from Backland et al. [12] and the new heart failure data. Consequently, the two Heart failure data sets were collapsed to one data set.

Participants

In order to recruit participating doctors for the Heart failure study, fifty general practitioners, selected randomly from 332 general practitioners in the southern part of Stockholm County, Sweden, and all 38 cardiologists at two cardiology clinics were invited to participate in the study. The general practitioners were contacted by telephone and 33 of them agreed to participate. After a single reminder letter, 27 of the general practitioners and 22 of the cardiologists had responded. Medical students from two courses in general practice were also invited to participate. Because they were in a situation of dependence on the department of family medicine, they were offered full anonymity in responding, and there were no reminders. Of the 82 students, 21 responded. No compensation was given to the participants, except for individual feedback on the result. Two medical students and two cardiologists had to be removed from the analysis because their responses had no variation. The participating doctors in the Heart failure study were therefore 19 medical students, 27 general practitioners, and 20 cardiologists. In the Hyperlipidemia study, 60 out of 90 general practitioners who worked in the southeast area of Stockholm, Sweden, were randomly selected. Out of these 60, 38 responded. Thus, the present paper is based on data from 64 participants in the Heart failure study and 38 participants in the on Hyperlipidemia study.

Procedure

In each study, 40 patient cases, from actual patients, were presented to the participating doctors. When collecting the merged data from the Heart failure study, the participating doctors were asked, on a

scale from 0-100%, to indicate the probability that the patient suffered from any degree of heart failure. In the Hyperlipidemia study, the participating doctors were asked, on a scale from 0-100%, to indicate their willingness to prescribe a lipid-lowering drug. The patient cases and the cue information in each case were presented in the same order to each doctor.

Material

In the Hyperlipidemia study, the patient cases contained seven information cues: age, sex, cholesterol value, triglyceride value, diabetes (yes/no), hypertension (yes/no), and history of coronary heart disease (yes/no). In the Heart failure study, the patient cases contained ten information cues: age, sex, history of myocardial infarction (yes/no), dyspnea (yes/no), atrial fibrillation (yes/no), leg edema (yes/no), rales (yes/no), systolic blood pressure (mm Hg), cardiac volume (ml/m^2), and signs of pulmonary congestion on lung X-ray (yes/no).

Coding of responses

In the Hyperlipidemia study, responses were dichotomized by considering the response as "Yes" (coded as 1) if the response was equal or greater than 50%, and considering the response as "No" (coded as 0) if it was 50% or below. In both data sets, missing data in the doctor decision and judgments were coded as 1, that is willingness to prescribe in the Hyperlipidemia study (1 missing response out of 1520 responses) and likelihood of a heart failure in the Heart failure study (11 missing responses out of 1080 responses). The reason for this data imputation was that it was assumed to be more plausible that the doctors would try to avoid false negative than false positive outcomes.

The continuous cue variables in the Hyperlipidemia study were dichotomized as follows: age above 65 years, cholesterol value at 6.5 mmol/l, and triglyceride value at 2.2 mmol/l. These levels were based on the guidelines of the European Heart societies for Hyperlipidemia treatment (1991). The continuous cue variables in the Heart failure study were dichotomized as follows: age above 65 years, systolic blood pressure below 140 mm Hg, and heart volume above 490 ml for men and 450 ml for women. The chosen levels of the different cues for the Heart failure study were mainly based on medical guidelines issued by Medical Products Agency in Sweden (1996).

Case sampling and model building

In order to randomize data selection for the fitting and prediction sets, and build models based on F&F and LR, Matlab version 2008 with statistics toolbox was used. For randomization of which cases to include in the fitting sets, we used Monte Carlo simulation (100 times for each sample and each model). The random function generated two sets of mutually exclusive sets of integers between 1 and 40. The first set of the numbers defined which of the total 40 cases to be included in the fitting sets for which the fit was calculated, the remaining set of the numbers defined which cases to be in the prediction set. If the randomly chosen cases for the fitting set consisted of all response values equal to zero or all equal to one, then this random set was excluded and a new one was generated. In order to be able to compare the impact of different sample sizes in cross-validation, four fitting set sizes; 25% of the total data set (10 cases), 37.5% of the total dataset (15 cases), 50% of the total dataset (20), and 75% of the total dataset (30 cases) were selected randomly and used for the modeling.

Model building - Fast and frugal model version 1 (F&F 1)

This model simulation was based on the F&F model described by Dhami and Harries [6] and conducted in a 4-step process: (1) the critical value of each cue was identified by choosing the value that had the highest number of positive responses (= positive diagnosis of heart failure or the decision to prescribe medical treatment). If the absolute frequency of the number of positive responses was the same for both values 1 and 0, the critical value was based on the lowest absolute frequency of negative responses. When these were also the same, the critical value was chosen randomly. Once the critical value was determined for each cue, (2) the validity value for each cue was calculated. The validity value tells us in which order the cues are searched by the model. The validity of a cue was calculated as the proportion of cases with the critical value for which a diagnosis or prescription of medical treatment was made. When the validity of all cues had been calculated (3) the validities were rank ordered in descending order. Cues with the same validity were ordered in the order they were presented to the doctors. Finally, (4) to determine the number of cues in the model, the percentage of correctly predicted decisions was first calculated for the cue with the highest validity. Afterwards, the second cue was added, and if the percentage of the fit was increased, this cue was also included in the model. This step was repeated until the percentage of correct predictions did not improve.

The number of cues searched before a decision was made was calculated for each doctor by applying the strategy to each case. If none of the k cues predicted a positive response (coded as 1) for a certain case, the response was assumed to be negative (coded as 0).

Model building - Fast and frugal model version 2 (F&F 2)

The simulation of F&F 2 was based on the model described by Backlund et al. [12]. This model was an extension of Dhami and Harries's [6] F&F model. The only difference between F&F 1 and F&F 2 was the identification of the critical value (step 1). In F&F 2, the critical value was based on the total number of correct predictions, that is, also including negative responses that were correct predictions. Thus, Dhami and Harries' model is based only on hits, whereas Backlund's model also includes correct rejections.

Model building - Logistic Regression (LR 5, LR 10, and LR Enter)

The independent variables in the LR models were the cues provided in the patient cases, eight cues from the Hyperlipidemia study and 10 cues from the Heart failure study. The dependent variable in the Hyperlipidemia study was the prescription decisions and in the Heart failure study, it was participant judgments and actual diagnoses. Stepwise forward with significance levels 5% (LR 5), 10% (LR 10), and Enter (LR Enter), corresponding to each of the three tested LR models, was used as the method for inclusion of cues. Stepwise forward calculates the residuals by adding a cue at a time in the model while Enter calculates the residuals for all the cues in the model. If no significant cues were found, only the constant was used for calculating the probability of a positive response. If the calculated probability of a positive response was equal to or greater than 50%, then the response was coded as a positive response (coded as 1); if it was smaller than 50%, then the response was coded as a negative response (coded as 0).

Data Analysis

Repeated measures ANOVA were conducted for testing the differences between the models with respect to predictive accuracy (proportion of correct predictions of each doctor's responses) and frugality. Frugality was measured on two dimensions; number of cues in model and number of cues utilized. Number of cues in a model concerns the optimal number of cues the model consists of in order to

provide the best predictive accuracy. Number of cues utilized in model concerns how many of the cues in the he model that is actually utilized for predicting the response in a given case. When there was a violation of sphericity in the model variances, Greenhouse-Geisser Correction was used for adjustment of sphericity. Post hoc tests of the five models (10 comparisons) were conducted using Holm's multistage Bonferroni. Comparison in terms of fit to the fitting sample was excluded because a model's predictive accuracy (i.e., ability to predict responses in the prediction set) gives more accurate information of how well the model describes the data.

In the data set from the Heart failure study, because the fitting set with 25% of the total data contained only ten cases and also the number of cues was ten, LR Enter analysis for this sample size was not conducted. In the frugality analysis, LR Enter was excluded in all cases because it uses all the cues.

Results

Predictive accuracy

Hyperlipidemia: A 5 (model) × 4 (fitting set sample size) within-subjects ANOVA on the predictive accuracy provided by the different models showed a significant main effect of model, F (2.56, 94.84) = 20.27, p<0.001. As can be seen in Figure 1, LR Enter provided the lowest predictive accuracy in all the fitting set sample sizes. Holm's adjusted Bonferroni revealed that regarding the different models the differences in predictive accuracy were significant between LR Enter and all other models (p<0.005). No other model differences were significant (p>0.05).

There was no main effect of sample size. There was, however, a significant interaction effect between model and sample size, F (3.89, 144.02)=15.62, p<0.001. The interaction effect is probably because the predictive accuracy of LR Enter decreased the smaller the fitting set sample size was. In order to control for this, the LR Enter values were removed and a 4 (model) × 4 (fitting set sample size) within-subjects ANOVA on the predictive accuracies was conducted. As expected, the interaction effects disappeared without the LR Enter model.

Heart Failure: Because the fitting set consisting of the 25% of all the data equaled 10 data-points, which was the same as number of cues in the LR Enter model, we conducted two ANOVAs, one excluding LR Enter while including fitting set 25% and one including LR Enter while excluding fitting set 25%. A 4(model) × 4 (fitting set sample size) within-subjects ANOVA on the predictive accuracy provided by the different models (excluding LR Enter but including fitting set 25%) showed no significant main effect of model, p>0.05. There was a significant main effect of fitting set sample size, F (1.12, 72.98)=14.01, p<0.001. Holm's adjusted Bonferroni revealed that regarding the different predictive accuracy provided by the different fitting set sample sizes, the differences were significant between fitting set size 37.5% and all other fitting set sizes (p<0.001), and between fitting set size 50% and all other fitting set sizes (p<0.005), in all cases with higher predictive accuracy in the greater fitting samples. There was also a significant interaction effect of model and sample size, F (2.77, 179.87)=13.00, p<0.001. As seen in Figure 2, the LR models, using 5% and 10% significance for cue inclusion did slightly better (1 or 2% higher accuracy) than the other models in all samples except for the fitting set consisting of 75%.

A 5 (model) × 3 (fitting set sample size) within-subjects ANOVA on predictive accuracy provided by the different models (including LR Enter but excluding fitting set 25%) showed a significant main effect of model, F (1.49, 97.13)=31.73, p<0.001. A seen in Figure 2, the LR models, using 5% and 10% significance for cue inclusion provided the highest predictive accuracy than the other models in all fitting set sizes while LR Enter provided the smallest predictive accuracy compared to the other models. Holm's adjusted Bonferroni revealed that regarding the different models differences were significant between LR Enter and all other models (p<0.001).

There was also a significant main effect of sample size, F (1.03, 66.63)=29.79, p<0.001, Holm's adjusted Bonferroni revealed that regarding the different fitting sizes, the differences were significant between all sample sizes (p<0.001).

Instead of the judgments provided by the doctors, the different fitting sets were also modeled against the actual diagnoses. As can be seen in Figure 3, when not cross-validated LR Enter provided the

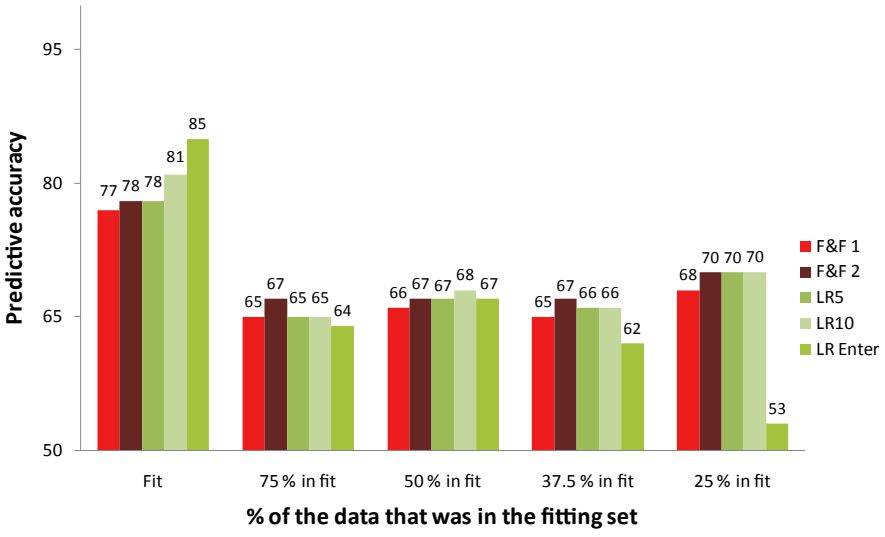

Figure 1: Percentage of mean fit and mean predictive accuracy provided by the different models and different fitting set sizes in the hyperlipidemia study.

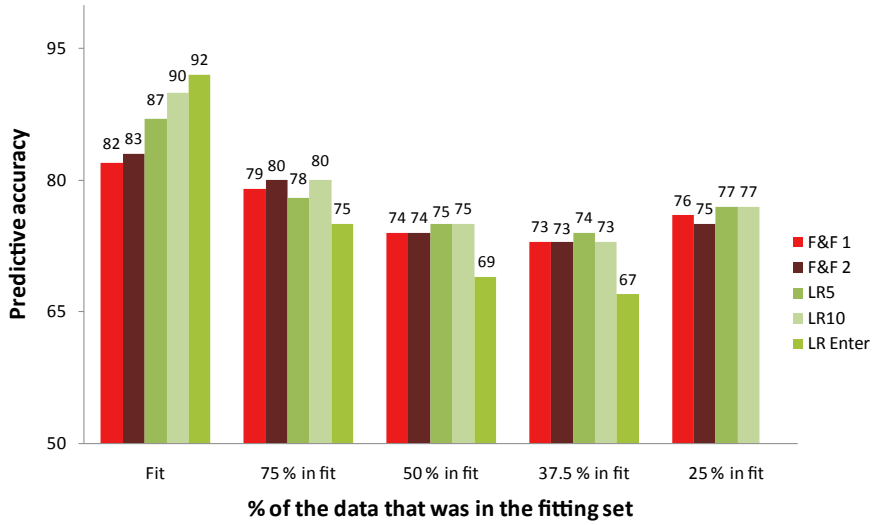

Figure 2: Percentage of mean fit and mean predictive accuracy provided by the different models and different fitting set sizes in the heart failure study.

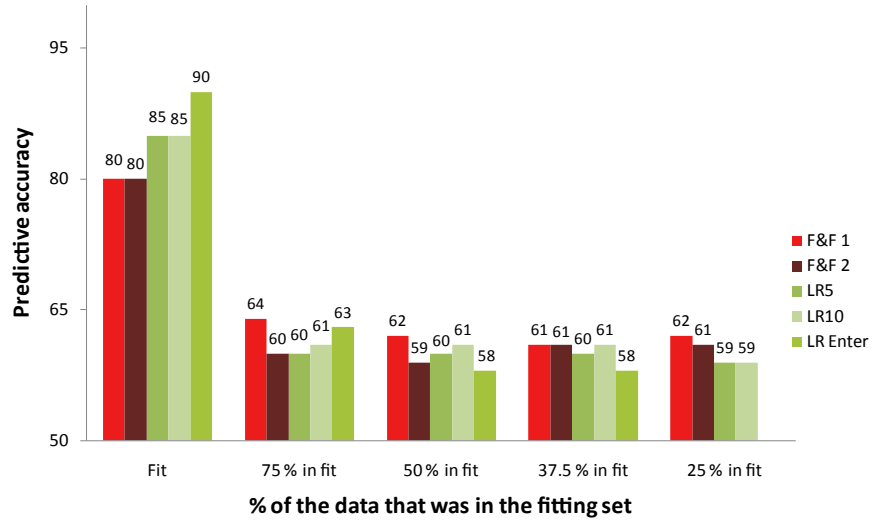

Figure 3: Percentage of mean fit and mean predictive accuracy provided by the different models and different fitting set sizes in the heart failure study, using actual diagnoses.

best fit but when cross-validated in all fitting sample sizes, the F&F 1 model did equally well or better than any other model, although the differences in predictive accuracy are small between different models, range being 0.58-0.63. Here, no statistical tests could be performed because N=1 (corresponding to one hypothetical doctor that would follow an optimal judgment model).

Frugality

Hyperlipidemia – Cues in Model: A 5 (model) × 4 (fitting set sample size) within-subjects ANOVA on the number of cues in each model showed a significant main effect of model, $F(1.05, 38.67) = 24.95$, $p<0.001$. As can be seen in Table 1, the LR 5 model consisted of fewer cues than any other model, which was true for all fitting sample sizes. Holm's adjusted Bonferroni revealed that regarding the different models, the differences were significant between all models, $p<0.001$, except for F&F 1 and LR 10.

Number of cues in the model					
	Fit	75% in fit	50% in fit	37.5% in fit	25% in fit
F&F 1	2.37	2.34	2.37	2.4	2.53
F&F 2	2.32	2.39	2.56	3.03	3.19
LR 5	1.74	1.76	1.56	1.45	1.18
LR 10	2.63	2.41	2.19	2.06	1.63
LR Enter	7	7	7	7	7
Number of cues actually utilized					
	Fit	75% in fit	50% in fit	37.5% in fit	25% in fit
F&F 1	1.61	1.69	1.27	1.61	1.31
F&F 2	1.68	1.77	1.36	2.05	2.1
LR 5	1.74	1.76	1.56	1.45	1.18
LR 10	2.63	2.41	2.19	2.06	1.63
LR Enter	7	7	7	7	7

Table 1: Frugality of the models and the different fitting set sizes in the hyperlipidemia study.

There was no significant main effect of the different fitting set sample sizes. There was, however, a significant interaction effect of sample size and model, F (2.07, 76.62)=60.97, p<0.001. As seen in Table 1, the smaller the fitting set sample size, the more cues in the model for the F&F models while the opposite is true for the LR 5 and LR 10 models; the smaller the fitting set sample size, the fewer cues in the model. It can be noted that for the greatest fitting samples, the mean number of cues in model was clearly lower for LR 5 (1.18, 1.45) than for F&F 1 (2.53, 2.40) and F&F 2 (3.19, 3.03).

Hyperlipidemia – Cues Utilized: A 5 (model) × 4 (fitting set sample size) within-subjects ANOVA on the number of cues in each model showed a significant main effect of model, F (1.11, 40.88) =28.59, p<0.001, where in the majority of the times, the F&F 1 model utilized the fewest number of cues (see Table 1) and LR Enter (by definition) utilized the largest number of cues, followed by LR 10 and F&F 2. Holm's adjusted Bonferroni revealed that the differences between models and number of cues utilized were significant in all cases (p<0.01) except between F&F 1 and LR 5 (p>0.05), although the differences were very small between F&F 2 , LR 5 and LR 10 when the fitting set was small (25%). There was also a significant main effect of the different fitting set sample sizes on number of cues utilized, F (1.43, 53.00)=31.27, p<0.001. Holm's adjusted Bonferroni revealed that the differences were significant between fitting set sample 25% and both 37.5% and 75% (p<0.001), between 37.5% and 50% (p<0.001), between 50% and 75% (p<0.001). There was also a significant interaction effect between sample size and model, on number of cues utilized F (1.88, 69.63)=41.68, p<0.001.

Heart Failure – Cues in Model: A 4(model) × 4(fitting set sample size) within-subjects ANOVA on the number of cues in each model showed a significant main effect of model, F (1.08, 70.38) =87.63, p<0.001, where the LR 5 model had the fewest cues in majority of the fitting sets while LR 10 had more cues than any other model (see Table 2). Holm's adjusted Bonferroni revealed that regarding the different models the differences were significant between all models (p<0.005) except for F&F 1 and LR 10.

There was also a significant main effect of fitting set sample size, F (1.27, 82.23)=104.05, p<0.001, where Holm's adjusted Bonferroni revealed that the differences were significant between all fitting set sample sizes (p<0.001). There was also a significant interaction effect between sample size and model, F (1.70, 110.22)=202.41, p<0.001. The number of cues in the F&F models increased with decreasing fitting set sample size, while the number of cues in the LR models decreased with decreasing fitting set size. Again it can be seen that for the smallest fitting set samples the mean number of cues in model was clearly lower for LR 5 (1.21, 1.58) than for F&F 1 (2.83, 2.65) and F&F 2 (3.18, 3.23).

As can be seen in Table 2 for actual diagnoses, in all fitting sets, LR 5 had the fewest numbers of cues in the model while F&F 2 model had more cues than any other model.

Heart Failure – Cues Utilized: A 4 (model) × 4 (fitting set sample size) within-subjects ANOVA on the number of cues utilized by each model showed a significant main effect of model, F (1.11, 72.26)=36.30, p<0.001, where the F&F 1 model utilized the fewest number of cues in half of the samples (Table 2). Holm's adjusted Bonferroni revealed that the differences were significant between all models (p<0.005). There was also a significant main effect of fitting set sample size on number of cues utilized, F (1.12, 73.03)=9.52, p<0.001. Holm's adjusted Bonferroni revealed that the differences were significant between fitting set sizes 75% and both 37.5% and 50%, between 37.5% and 75%, and 50% and 75%, p<0.05. There was also a significant interaction effect between model and sample size, F (1.93, 125.16)=99.73, p<0.001. The number of cues in F&F 2 increased in the smaller fitting sets.

Number of cues in the model					
	Fit	75% in fit	50% in fit	37.5% in fit	25% in fit
Participant judgments					
F&F 1	2.07	2.39	2.53	2.65	2.83
F&F 2	2.93	2.95	2.99	3.23	3.18
LR 5	2.48	2.43	1.9	1.58	1.21
LR 10	3.33	3.44	2.85	2.34	1.73
LR Enter	10	10	10	10	10
Actual diagnoses					
F&F 1	3	2.73	2.56	2.59	2.75
F&F 2	4	3.19	2.96	3.12	3.15
LR 5	2	1.65	1.78	1.41	1.19
LR 10	2	2.36	2.48	2.34	1.77
LR Enter	10	10	10	10	10
Number of cues actually utilized					
	Fit	75% in fit	50% in fit	37.5% in fit	25% in fit
Participant judgments					
F&F 1	1.55	1.39	1.78	1.58	1.72
F&F 2	1.99	1.56	1.35	1.78	2.01
LR 5	2.48	2.43	1.9	1.58	1.21
LR 10	3.33	3.44	2.85	2.34	1.73
LR Enter	10	10	10	10	10
Actual diagnoses					
F&F 1	2.36	2.1	1.97	2	2.05
F&F 2	2.88	2.35	1.49	2.34	2.33
LR 5	2	1.65	1.78	1.41	1.19
LR 10	2	2.36	2.48	2.34	1.77
LR Enter	10	10	10	10	10

Table 2: Frugality of the models and the different fitting set sizes in the heart failure study.

Turning to actual diagnoses, as can be seen in Table 2, in all fitting sets, LR 5 utilized fewest numbers of cues, followed by F&F 1 while in majority of the times F&F 2 utilized most number of cues. Similar to the judgment data, in the LR models, the smaller fitting set size, the smaller number of cues in the model. Regarding number of cues in model, one-sample t-tests showed that in sample size 75% both F&F 1 (t (65)=-4.34, p<.001) and F&F 2 (t (65) = -3.11, p<.005) used more cues in actual diagnoses than in participant judgments.

One-sample t-tests showed that for all sample sizes both F&F models used more cues in predictions of actual diagnoses as compared to predictions of participant judgments (p<.001 in all 8 comparisons), whereas the reverse was true for LR 5 and LR 10, which tended to utilize fewer cues in predictions of actual diagnoses than in predictions of participant judgments across different sample sizes with significant results for sample 75% , LR 5 (p<.001) and LR 10 (p<.001), for sample 50%, LR 5 (p<.05) and LR 10 (p<.001), for sample 37.5% , LR 5 (p<.001), and for sample 25% (p<.001).

Discussion

Previously, most of the model comparisons involving F&F and LR models have modeled ecological relations rather than how people use ecological information in judgments or decisions [7]. In studies where human judgments or decisions were modeled, cross-validation was not used for comparing the validity of F&F and LR models [6],

as far as we know. In studies where cross-validation was used, the LR models have used all the cues in the model [15]. Using all the cues in the model leads to an artificial lowering of predictive accuracy after cross-validation because the LR-model is over-fitted to the original data. F&F models may be assumed to be less sensitive to over-fitting, because they are constructed to be fast decisions based on a minimum of information. For this reason, the comparisons between LR and F&F models in previous studies using cross-validation have been biased against LR models. In the present study, two F&F models and three LR models with different cue inclusion levels were compared to each other on different fitting sets consisting of both ecological data (actual diagnoses) and judgment data (participant judgments). In addition, the models' frugality was compared on two dimensions.

It may be argued that the actual diagnoses, which were treated as ecological data in this study, in fact involve judgments and therefore are not pure ecological data. However, the actual diagnoses are as close as we can come to the ecological fact of the patients' true diagnosis. Thus, we regard the actual diagnoses as a proxy for the true diagnoses.

Predictive accuracy, sample size and inclusion level

In both studies, LR Enter had the highest fit, but when cross-validated it had the lowest predictive accuracy. There were no significant differences in predictive accuracy between the F&F and the logistic regression models (not including LR Enter) in the Hyperlipidemia study. By contrast, in the Heart failure study there was a slightly higher predictive accuracy for the other two LR models compared to each of the F&F models. In the Hyperlipidemia and Heart failure study, the smaller the fitting set sample size, the lower predictive accuracy of LR Enter. In both studies, because the predictive accuracy of LR Enter decreased more compared to all other models, it is plausible to assume that inclusion level plays an important role when evaluating the predictive accuracy of an LR-model. Apparently, this was not checked in previous studies where only LR Enter was compared with F&F models.

Regarding sample size and predictive accuracy, in the Hyperlipidemia study, the predictive accuracy was unrelated to size of the fitting sample, except for LR Enter. In the Heart failure study, the different sample sizes yielded different levels of predictive accuracy. In particular, and, in line with the results from Martignon et al., it seems that the greater the fitting sample, the higher predictive accuracy in the LR Enter model. This emphasizes the argument that the comparisons of LR *vs.* F&F models have not been fair because the only regression model that has been tested is LR Enter. Allowing a higher significance level for cue inclusions gives more capitalization on chance [18] because there is a greater risk that irrelevant cues are included in the model. This capitalization should reveal itself when a fitted LR Enter model is used for calculating the predictive accuracy because the fitted model is not performance maximized for the prediction set. In fact, this was the case when the data were cross-validated, where the predictive accuracy of LR Enter decreased markedly and in most cases yielded less accurate predictions than the other models. In their regression model, Dhami and Harries used a significant level of 5% and they found that both the F&F and the regression model gave equally good fit. Backlund et al. used a significance level of 10% and found that the regression model provided better fit. When it comes to predictive accuracy, the choice of 5% *vs.* 10% seems to have little importance than is true for the choice of 5% or 10% *vs.* 100% (LR Enter).

To conclude, there should be a greater emphasis in testing regression models with different significance levels of cue inclusion in the models. As showed in the present study, the lower significance level of cue inclusion gave different results than the one including all the cues. These differences might be one factor that explains the seemingly contradictory results in judgment studies as compared to ecological studies.

Frugality

The different models were also compared in terms of frugality. In both the Hyperlipidemia and the Heart failure study, in the majority of the times, LR 5 included fewer cues in model, while LR 10 and obviously LR Enter included the more cues in model as compared to the F&F models. Because the F&F models, in contrast to LR models, are designed to use a minimum of information, one would expect the LR models to use more cues than the tested F&F models. However, this was not the case. When only significant cues are considered, the LR models, especially LR 5, tended to be more frugal as compared to the F&F models.

Previous studies, have measured frugality in terms of number of cues in a model. However, a model might not necessarily utilize all the cues is consists of. It is equally important to investigate how many cues a model actually utilizes because heuristics may use fewer cues in order to implement non-exhaustive search. Regarding the number of cues actually utilized by the F&F models, the F&F 1 model utilized fewer numbers of cues compared to F&F 2, however, not consistently fewer than LR 5. Important here is also the difference between the numbers of cues a model consists of and number of cues actually utilized by the model. The smaller difference, the better is the model in the sense that it actually makes use of the cues the model consists of. If a model consists of, say, 3 cues but only utilizes 1 of these cues, then two of these cues are superfluous and as a result the number of cues is not "saturated". In the LR models, the same number of cues is used both in model and in utilization, meaning that they are fully saturated. In both Hyperlipidemia and the Heart failure study, for both F&F models, there was a difference varying from 26 to 48% less cues utilized than total number of cues available in the model. However, this was dependent on the size of the fitting sample. The smaller the fitting sample, the greater was this difference.

Actual diagnoses *vs.* participant judgments

When the data were not cross-validated, both the actual diagnoses and participant judgments yielded the same predictive accuracy in all the models. However, things were different when the data were cross-validated, where predictive accuracy was lower for actual diagnoses than for participant judgments. In terms of frugality, the F&F models tended to use more cues in actual diagnoses than in participant judgments, especially with respect to number of cues actually utilized. The opposite was true for the LR models inasmuch as LR 5 and LR 10 used less cues in model (which for LR models coincides with number of cues actually utilized) in the actual diagnoses as compared to the participant judgments. These differences combined with the relatively poor fit found in actual diagnoses after cross-validation suggest that other factors than the ones investigated in this study were important for the actual diagnoses. On the other hand, the predictive accuracy for actual diagnoses was approximately the same for all models except LR Enter as also was true for the corresponding participant diagnoses. This finding adds to the validity of the outcome of the model comparisons made in the present study.

Conclusions

As a whole, our data suggest that also when cross-validation is used, linear models behave as efficiently, if not more efficiently, than fast and

frugal models, provided that only significant cues are included in the model. This is in line with recent research using experimental methodology showing that compensatory judgment models (e.g., linear models) may outperform F&F models [15,22-24] when applied to behavioral judgment data or that the validity of F&F models may have been overestimated [25].

Actually, also the linear models tested in the present study deserve to be called fast and frugal. The mean number of cues in the fitted LR models was mostly only one or two cues selected from a total of 7 (Hyperlipidemia study) or ten cues (Heart failure study). Thus, only a small proportion of the given information was utilized in the LR models, but perhaps in a more nuanced way than in F&F models, depending on the possibility to model compensatory judgments. This may explain why the predictive accuracy tended to be slightly higher for LR 5 and LR 10 as compared to for the two F&F models.

As a final note of caution, the data in the present study are based on means calculated over all the participating doctors and the models are built using the means. It could be the case that some individuals may be consistently using a heuristic as found in Skånér et al. [26]. In fact they might even be using heuristics other than those investigated in this study, which might predict judgments and decisions better than the strategies investigated in the present study. This could be an interesting topic for future research but so far our data suggest that LR models could be as fast and frugal than the original F&F models [27-30].

The take-home message of this study is that in both actual and behavioral judgments in the medical domain as well as in other fields cross-validation coupled with a systematic variation of significance levels for cue selection is important when logistic regression is compared with other judgment models. The cross-validation showed, contrary to some previous results based on model fitting rather than cross-validation, that LR is equally good, if not better, in capturing human decision making as F&F models and depending on cue inclusion level LR models might be more frugal than F&F models.

References

1. Engel JD, Wigton RS, LaDuca A, Blacklow RS (1990) A social judgment theory perspective on clinical problem solving. Evaluation and the Health Professions 1: 63-78.
2. Wigton RS (1996) Social judgment theory and medical judgment. Thinking and Reasoning 175-190.
3. Hamm R, Yang H (2016) Alternative lens model equations for dichotomous criteria, Journal of Behavioral Decision Making. Wiley Online Library.
4. Gigerenzer G (2015) Simply rational. Decision making in the real world. Oxford: Oxford University Press.
5. Gigerenzer G, Goldstein DG (1999) Betting on one good reason: The take the best heuristic. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, Simple heuristics that make us smart. New York: Oxford University Press 975-995.
6. Dhami M, Harries C (2001) Fast and frugal versus regression models of human judgment. Thinking and Reasoning 7: 5-27.
7. Gigerenzer G, Goldstein DG (1996) Reasoning the fast and frugal way: models of bounded rationality. Psychol Rev 103: 650-669.
8. Gigerenzer G, Todd PM (1999) Fast and frugal heuristics: the adaptive toolbox. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), Simple heuristics that make us smart. New York: Oxford University Press 3-34.
9. Gigerenzer G, Hoffrage U, Kleinbölting H (1991) Probabilistic mental models: a Brunswikian theory of confidence. Psychol Rev 98: 506-528.
10. Marewski JN, Gigerenzer G (2012) Heuristic decision making in medicine. Dialogues Clin Neurosci 14: 77-89.
11. Graefe A, Armstrong JS (2012) Predicting elections from the most important issue: A test of the take-the-best heuristic. Journal of Behavioral Decision Making 2: 41-48.
12. Backlund L, Bring J, Skånér Y, Strender LE, Montgomery H (2009) Improving fast and frugal modeling in relation to regression analysis: Test of three models for medical decision making. Medical Decision Making 29: 140-148.
13. Czerlinski J, Gigerenzer G, Goldstein DG (1999) How good are simple heuristics? In G. Gigerenzer, P. M. Todd, & the ABC Research Group, Simple heuristics that make us smart (pp. 97–118). New York: Oxford University Press.
14. Kee F, Jenkins J, McIlwaine S, Patterson C, Harper S, et al. (2003) Fast and frugal models of clinical judgment in novice and expert physicians. Medical Decision Making 2: 293-300.
15. Martignon L, Katsikopoulos KW, Woike JW (2008) Categorization with limited resources: A family of simple heuristics. Journal of Mathematical Psychology: 352-361.
16. Smith L, Gilhooly K (2006) Regression versus fast and frugal models of decision making: The case of prescribing for depression. Applied Cognitive Psychology 20: 265-274.
17. Plutowski M, Sakata S, White H (1994) Cross-validation estimates IMSE. in Cowan, J.D., Tesauro, G., and Alspector, J. (Eds.) Advances in Neural Information Processing Systems 6 (pp. 391-398), San Mateo, CA: Morgan Kaufman.
18. Pitt MA, Myung IJ, Zhang S (2002) Toward a method of selecting among computational models of cognition. Psychological Review 109: 472-491.
19. Martignon L (2001) Comparing fast and frugal heuristics and optimal models. In G. Gigerenzer & R. Selten (Eds.), Bounded rationality: The adaptive toolbox. Cambridge, MA: MIT Press 147-171.
20. Martignon L, Hoffrage U (2002) Fast, frugal, and fit: Simple heuristics for paired comparison. Theory and Decision 29-71.
21. Dhami MK (2003) Psychological models of professional decision making. Psychol Sci 14: 175-180.
22. Ayal S, Hochman G (2009) Ignorance or integration: The cognitive processes underlying choice behavior. Journal of Behavioral Decision Making 2: 455-474.
23. Glöckner A, Betsch T (2008) Multiple-reason decision making based on automatic processing. Journal of Experimental Psychology: Learning, Memory, and Cognition 3: 1055-1075.
24. Su Y, Rao LL, Sun HY, Du XL, Li X, et al. (2013) Is making a risky choice based on a weighting and adding process? An eye-tracking investigation. Journal of Experimental Psychology: Learning, Memory, and Cognition 39: 1765-1780.
25. Jekel M, Glöckner A (2016) How to identify strategy use and adaptive strategy selection: The crucial role of chance correction in weighted compensatory strategies. J Behavioral Decision Making. Wiley Online Library.
26. Skaner Y, Bring J, Ullman B, Strender LE (2000) The use of clinical information in diagnosing chronic heart failure: A comparison between general practitioners, cardiologists, and students. J ClinEpidemiol 53: 1081-1088.
27. Svenson O (1979) Process descriptions of decision making, Organizational Behavior and Human Performance 2: 86-112.
28. Wigton RS (1996) Social judgment theory and medical judgment. Thinking and Reasoning 175-190.
29. Dhami MK, Schlottmann A, Waldmann MR. (2012) Judgment and decision making as a skill: Learning, development and evolution, by Dhami, Mandeep K.; Schlottmann, Anne; Waldmann, Michael R. Cambridge University Press 291-306.
30. Dhami MK, Ayton P (2001) Bailing and jailing the fast and frugal way. Journal of Behavioral Decision Making 1: 141-168.