Optimal threshold estimator of a prognostic marker by maximizing a time-dependent expected utility function for a patient-centered stratified medicine

Etienne Dantan, Yohann Foucher, Marine Lorent, Magali Giral, Philippe Tessier

Optimal threshold estimator of a prognostic marker by maximizing a time-dependent expected utility function for a patient-centered stratified medicine

Etienne Dantan[1], PhD, Yohann Foucher[1,2], PhD, Marine Lorent[1], PhD, Magali Giral[3,4], MD, PhD, Philippe Tessier[1], PhD

[1] EA 4275 SPHERE – methodS in Patient-centered outcomes & HEalth ResEarch. Nantes University, France.

[2] Nantes University Hospital, Nantes, France.

[3] Transplantation, Urology and Nephrology Institute (ITUN), CHU Nantes, RTRS « Centaure », Nantes and Inserm U1064, Labex Transplantex, Nantes University, France.

[4] Centre d'Investigation Clinique biothérapie, 30 bd Jean Monnet, 44093, Nantes, France.

Short title: Threshold estimation of a prognostic marker

Corresponding author:

Etienne Dantan, EA 4275 SPHERE – methodS in Patient-centered outcomes & HEalth ResEarch, Nantes University, 22 boulevard Bénoni Goullin, 44200 Nantes, France.

Telephone: 33 2 53 00 91 23; Fax: 33 2 40 41 29 96

e-mail: Etienne.Dantan@univ-nantes.fr

Abbreviations

AUC, Area Under the Curve; CI, Confidence Interval; DCA, Decision Curve Analysis; KTFS, Kidney Transplant Failure Score; QALYs, Quality-Adjusted Life-Years; QoL, Quality of Life; RMST, Restricted Mean Survival Time; ROC curve, Receiver Operating Characteristic curve; SE, Standard Error

Abstract

Defining thresholds of prognostic markers is essential for stratified medicine. Such thresholds are mostly estimated from purely statistical measures regardless of patient preferences potentially leading to unacceptable medical decisions. Quality-Adjusted Life-Years (QALYs) are a widely used preferences-based measure of health outcomes. We develop a time-dependent QALYs-based expected utility function for censored data that should be maximised to estimate an optimal threshold. We performed a simulation study to compare estimated thresholds when using the proposed expected utility approach and purely statistical estimators. Two applications illustrate the usefulness of the proposed methodology which was implemented in the R package *ROCt* (www.divat.fr). Firstly, by reanalysing data of a randomized clinical trial comparing the efficacy of prednisone versus placebo in patients with chronic liver cirrhosis, we demonstrate the utility of treating patients with a prothrombin level higher than 89%. Secondly, we reanalyse the data of an observational cohort of kidney transplant recipients: we conclude to the uselessness of the Kidney Transplant Failure Score (KTFS) to adapt the frequency of clinical visits. Applying such a patient-centered methodology may improve future transfer of novel prognostic scoring systems or markers in clinical practice.

Keywords: prognostic marker, threshold definition, QALY, censored data, stratified medicine, patient preferences

Introduction

Stratified medicine currently relies on the use of biological markers capable of discriminating patients into sub-groups receiving the most appropriate treatment (1,2). The recent opportunities towards this personalized patient care, especially offered by omics technologies, have transformed biomedical research (3). Despite this, a very low proportion of discovered biomarkers or scoring systems have been transferred to clinical practice. This disillusion is partially explained by the numerous methodological pitfalls, leading to low credibility of the results (4,5), and by the non-consideration of the patients themselves: studies are often isolated from their social or environmental context (6), and independent of their preferences over the possible outcomes of individual-tailored clinical choices. In the literature, the potential usefulness of prognostic markers is often judged regarding purely statistical measures such as the highest area under the ROC curve (AUC). This indicator is widely associated with the following normative premise: the higher the AUC, the lower the error rates of a decision based on the marker, and the higher the corresponding usefulness in future medical decision making. Several common approaches (7,8) can be envisaged to estimate a marker threshold in order to discriminate between low and high risk patients, as the Youden index corresponding to the marker value that maximizes the sum of sensitivity and specificity.

However, as recently recalled by Subtil et al. (9), these purely statistical approaches had already been clearly criticized in the literature since they ignored patient preferences of the health outcomes consequences of the stratified medical decision making, possibly leading to unacceptable trade-offs between the harms and the benefits of a clinical decision (10,11). From a decision analysis framework, several theoretical-based papers have shown how a clinically

useful threshold for a diagnostic test can be determined by maximizing an expected utility function, utility referring to the intensity of individual preferences over health outcomes (9,12–15). Even if the expected utility maximization in order to estimate optimal threshold on a continuous marker appears relatively well-known in diagnostic context, these decision analytic approaches remain difficult to extrapolate to time-dependent prognostic frameworks that imply dealing with right censoring data and time-dependent utility measures for health outcomes. Fortunately, the former point can be treated by considering survival models and the latter point can be handled using Quality-Adjusted Life Years (QALYs). QALYs are a widely used composite measure of the health care consequences that combines in a single number information about the quantity and the health-related Quality of Life (QoL) (16). Although QALYs have been primarily designed for the purpose of economic evaluation, their potential usefulness for clinical decision-making has been acknowledged and warrants further consideration (17–19). More precisely, QALYs represent any sequence of health states over time as an equivalent number of years lived in perfect health. Each life year is weighted by preference-based measures, called utility scores. The utility scores usually range from 0 (dead) to 1 (perfect health), so that a higher score indicates a more preferred health state. For instance, a patient alive 10 years with a utility of 0.8 will have 8 QALYs (10 * 0.8), a value higher than the one for a patient alive 12 years with a utility of 0.6, this last one having 7.2 QALYs (12 * 0.6) due to a more efficient intervention but with more side effects. The utility scores can be assessed directly by asking patients. For this purpose, different methods exist, such as the standard gamble or the time trade-off methods for instance (20). The utility scores can also be obtained indirectly from pre-scored generic health-states description systems, such as the Euroqol EQ-5D (21) or the SF-6D (22), that represent the preferences of the

general population. Another possibility is to gather utility scores through literature review (23). Given the widespread use of QALYs, published scores are available for a wide variety of health conditions.

The aim of this paper is to extend expected utility maximization to prognostic frameworks in order to estimate prognostic marker thresholds to achieve patient-centered stratified medicine. We performed a simulation study to highlight the interest of considering the individual preferences comparatively to purely statistical approaches. We illustrate the usefulness of the methodology by assessing two different clinical decision contexts. The first concerns reanalysis of an old clinical trial aimed at assessing prednisone prescription in patients suffering of chronic liver cirrhosis based on the prothrombin level (24). The second relies on an observational cohort aiming to use the Kidney Transplant Failure Score (KTFS) to adapt the follow-up of kidney transplant recipients after one year post-transplantation (25).

Methods

Definition of the time-dependent expected utility function for a stratified medical decision

Let a sample be constituted by n patients. Let T, C, Z and X be, respectively, the time to the time-to-failure, the time-to-censoring, the treatment (A or B) and the baseline prognostic marker under investigation to potentially drive the treatment allocation. For each subject i $(i = 1, \dots, n)$, we observe $\{y_i, z_i, x_i\}$ with $Y = \min(T, C)$. Let $D(\tau)$ be the indicator of failure such that $D(\tau) = 1$ if it occurs before the forecast horizon time τ, and $D(\tau) = 0$ otherwise. By convention, assume that the incremental benefit of the treatment A compared to the treatment B in terms of delayed failure increases with X at the price of possible side effects deteriorating QoL (again compared to

the treatment B). The aim is to define a threshold κ for the following decision: proposition of the treatment A for a patient i when $\{x_i > \kappa\}$ and B otherwise. Patient profiles may therefore be distinguished by combining their two possible initial strata ($X > \kappa$ or $X \leq \kappa$) and their two possible outcomes ($D(\tau) = 1$ or $D(\tau) = 0$). The stratified medical decision leads to four possible health outcomes that each correspond to specific flows of time-dependent health states or health-related QoL levels over the prognostic period. Adapting the initial proposition of Metz (13), the expected utility function correspond to a sum over all four health outcomes weighted by their corresponding utility. In a prognostic framework, taking into account the patients' preferences over health states, the optimal threshold κ^* can be estimated by maximizing the following time-dependent expected utility function:

$$\psi_\tau(\kappa) = \sum_{g \in \{X>\kappa, X \leq \kappa\}} \sum_{j \in \{0,1\}} P(g, D(\tau) = j) Q(\tau | g, D(\tau) = j) \qquad (1)$$

where $Q(\tau | g, D(\tau) = j)$ represent the numbers of QALYs up to the prognostic time τ corresponding to each possible profile and $P(g, D(\tau) = j)$ are the associated probabilities. The number of QALYs is defined as the sum of life years weighted by the instantaneous utility ranging from 0 (being dead) to 1 (being in perfect health) of the corresponding health states (26), i.e. a mean of the health-state utilities weighted by the corresponding probabilities up to time τ:

$$Q(\tau | g, D(\tau) = j) = \sum_{l \in \{0,1\}} \int_0^\tau u(t | g, D(t) = l) P(D(t) = l | g, D(\tau) = j) dt \qquad (2)$$

where $u(t | g, D(t) = l)$ is the instantaneous utility defined on the scale 0-1 (death-perfect health) of the health state at time t. We assumed, as the QALY model does, that the instantaneous utility scores at time t for health states are constant over time, i.e. $u(t | g, D(t) = l) = u_{g,l}$ for all t. If

$g = \{X > \kappa\}$, the utility of receiving treatment A is considered ($u_{X>\kappa,l} = u_{A,l}$); otherwise the utility of receiving B ($u_{X\leq\kappa,l} = u_{B,l}$). It can be shown that the time-dependent expected utility function (equation (1)) may then be simplified as follows (see demonstration in appendix):

$$\psi_\tau(\kappa) = \sum_{g\in\{X>\kappa, X\leq\kappa\}} P(g)\left[u_{g,0}E(\min(T,\tau)\,|g) + u_{g,1}\left(\tau - E(\min(T,\tau)\,|g)\right)\right] \qquad (3)$$

where $E(\min(T,\tau)\,|g)$ is the Restricted Mean Survival Time (RMST) up to time τ in the group g, i.e. the average survival time when patients are followed up to τ (27). Because we assumed that, compared to the treatment B, the treatment A is associated to a better efficacy in delaying the failure for high values of the marker X and a lower QoL due to side-effects, we also have to respect the condition $\left\{u_{X>\kappa,l} \leq u_{X\leq\kappa,l}\right\}$ for all $l \in \{0,1\}$. Because we considered a time-to-failure, we also have $\left\{u_{g,0} > u_{g,1}\right\}$ for all $g \in \{X > \kappa, X \leq \kappa\}$.

For a patient-centered medical decision, the time τ should be determined as the most important value with sufficient at-risk patients to allow a reliable statistical inference. Patient expectations are both to delay the event as late as possible and to maintain their QoL as long as possible. Note also that the value of τ enters in the equation (3), therefore the choice of τ affects the corresponding number of QALYs, resulting in the non-consideration of the decision consequences beyond τ. We therefore recommend to define τ as the highest possible value.

Note that the maximization of the time-dependent expected utility function (equation (3)) may result in an estimation of κ^* equal to an extremum of the marker. This implies treating all patients with A if $\kappa^* = \min(X)$, or with B if $\kappa^* = \max(X)$. These situations indicate the futility of the prognostic marker for patient-centered stratified medical decision making, even if its prognostic capacities are important.

Non-parametric estimation of the time-dependent expected utility function

We used a non-parametric approach in order to maximize the expected utility function leading to the optimal threshold κ^*. As demonstrated, the expected utility function can be decomposed into three terms (equation (3)). First, the probabilities $P(g)$ for $g \in \{X > \kappa, X \leq \kappa\}$, can be estimated by the empirical cumulative distribution function of X in g, i.e $P(g) = n^{-1} \sum_{i=1}^{n} 1(X_i \in g)$, with $1(\cdot)$ the indicator function equal to 1 if the condition within the brackets is verified and 0 otherwise. Second, the utility scores $u_{g,l}$, for $g \in \{X > \kappa, X \leq \kappa\}$ and $l \in \{0,1\}$, can be obtained from individual preference scores already published in the literature. Third, $E(\min(T, \tau) | g)$ can be estimated by the area under the Kaplan-Meier survival curve in patients of the strata g with the adequate treatment, noticed $\hat{S}(\cdot | g)$, up to time τ. More precisely, among patients of the strata g, let $t_{1,g} < \cdots < t_{q,g} < \cdots < t_{p,g}$ be the different event time, $d_{q,g}$ the number of subjects experiencing the event at time $t_{q,g}$ and $R_{q,g}$ the number of at-risk patients at the same time. Corresponding to the average survival time when patients are followed up to τ, the RMST can therefore be written as:

$$
\begin{aligned}
E(\min(T, \tau) | g) &= \int_0^{\tau} S(t|g)\, dt \\
&= \sum_{m: t_{m,g} \leq \tau} \left(t_{m,g} - t_{m-1,g}\right) \times \hat{S}\left(t_{m,g} | g\right) \\
&= \sum_{m: t_{m,g} \leq \tau} \left[\left(t_{m,g} - t_{m-1,g}\right) \times \prod_{q: t_{q,g} \leq t_{m,g}} \left(1 - \frac{d_{q,g}}{R_{q,g}}\right) \right]
\end{aligned}
$$

Depending on the study design, the RMST may be directly estimated in each strata g. For instance, in a clinical trial, the two treatments are observed: $E(\min(T,\tau)\,|X>\kappa)$ is then estimated in patients receiving treatment A, i.e. $\mathrm{E}(\min(\mathrm{T},\tau)\,|X>\kappa, Z=A)$, while $E(\min(T,\tau)\,|X\leq\kappa)$ is estimated in patients receiving treatment B, $\mathrm{E}(\min(\mathrm{T},\tau)\,|X\leq\kappa, Z=B)$. However, in some other cases, specific assumptions have to be formulated. For instance, from an observational cohort in which only treatment B is observed, $E(\min(T,\tau)\,|X>\kappa)$ cannot be directly estimated since treatment A is not observed. In this situation, we have to assume a relative increase (Δ) in the RMST of patients treated by A compared to B, i.e. $E(\min(T,\tau)\,|X>\kappa)=\min(E(\min(T,\tau)\,|X>\kappa, Z=B)\times(1+\Delta);\tau)$.

The 95% confidence interval (CI) of the optimal threshold κ^* can be estimated by using non-parametric bootstrap resampling (28). From the observational data, 2000 independent bootstrap samples are generated and the expected utility function is maximized in order to estimate an optimal threshold for each sample. The corresponding 2.5th and the 97.5th percentiles represent the 95% CI. All analyses were performed using the 3.0.2. version of the R software (29). This method has been implemented in the package *ROCt* available at www.divat.fr.

Simulation study

Design

Data were simulated according a 1:1 randomized clinical trial design. For each subject i $(i=1,\ldots,1000)$, the prognostic marker X_i was obtained from a Gaussian distribution truncated between 20 and 600 of mean 140 and standard deviation 55. The binary variable Z_i related to the treatment A $(Z=1)$ or B $(Z=0)$ was simulated according to a Bernouilli distribution with

probability at 0.5. The time-to-censoring C_i was generated from a Weibull distribution in order to obtain around 10% of censoring at 6 years. The time-to-event T_i was simulated from a proportional hazard model: $\lambda_0(t_i)\, exp(\beta_X X_i + \beta_Z Z_i + \beta_{XZ} X_i Z_i)$, with a Weibull distribution for the baseline risk $\lambda_0(t)$. Weibull parameters were chosen in order to have roughly a 30% survival probability at 6 years of follow-up. The prognostic marker X_i was assumed to be a risk factor with $\beta_X = 0.002$.

We considered different scenarios for which we simulated 1000 samples of patients. More precisely, we investigated three treatment effects $\beta_Z = (0, -0.5, -1.25)$ and three interaction levels with the marker $\beta_{XZ} = (0, -0.004, -0.008)$. Moreover, different impacts on QoL were investigated: a similar utility between the two treatments ($u_{A,0} = u_{B,0} = 0.7$), a 15% decrease due to treatment A compared to B ($u_{A,0} = 0.6;\ u_{B,0} = 0.7$), a 30% decrease due to treatment A compared to B ($u_{A,0} = 0.5;\ u_{B,0} = 0.7$). We considered the death as the event of interest, we therefore have $u_{A,1} = u_{B,1} = 0$. The prognostic time τ was fixed at 6 years.

We compared three estimators: our proposed approach based on the expected utility maximization, the Youden index, and the profile likelihood maximization of a Cox model including the treatment, the marker and the corresponding interaction.

Results

Simulation results of the different scenarios are presented in Table 3. The two purely statistical approaches approximately estimated a threshold round the value 140, i.e. the mean of prognostic marker. As expected, the Youden index and the profile likelihood maximization lead to thresholds

insensitive to the assumptions made on the utilities. It may result in unacceptable stratified medical decision making. For instance, when considering a higher efficacy of treatment A compared to treatment B ($\beta_Z = -1.25$) with no interaction with the marker ($\beta_{XZ} = 0$) and similar utilities ($u_{A,0} = u_{B,0} = 0.7$), the Youden index led to a threshold at 142.91 and the profile likelihood maximization at 141.88. It leads to recommend treatment B for about one half of the population, a nonsense regarding such an obvious balance benefit/risk in favour of treatment A. In this situation, maximizing the time-dependent expected utility leads to a threshold at 25.92 for which more than 99% of patients would receive treatment A. This threshold increased to 91.61 when utility for treatment A is 0.6 and utility for treatment B is 0.7, corresponding to treat 20% of patients with treatment A. In contrast, when side effects due to treatment A were too strong compared to treatment B $\left(u_{A,0} = 0.5;\ u_{B,0} = 0.7\right)$, the threshold was 281.49, corresponding maintain all patients with treatment B.

When the interaction coefficient is negative, the effect of treatment A increases with the marker and should lead to recommend treatment A to more patients. This trend was observed from our QALYs-based expected utility maximization. For example, for a 15% utility decrease $\left(u_{A,0} = 0.6;\ u_{B,0} = 0.7\right)$ and an intermediate treatment effect ($\beta_Z = -0.5$), the threshold was 126.48 for a $\beta_{XZ} = -0.004$ and 85.57 for a $\beta_{XZ} = -0.008$. In contrast, purely statistical approaches were not sensitive.

Considering obvious balances benefit/risk, such as the better efficacy of treatment A without more side effects compared to treatment B, or the similar efficacy of treatment A but with more side effects compared to the treatment B. In these situation, our proposed approach give

standard errors (SE) lower than the ones obtained from purely statistical approaches. In contrast, when the balances benefit/risk require an arbitration, SE obtained from our approach are higher than the SE obtained from purely statistical approaches. Therefore, it appears that the SE obtained from our approach better reflects the uncertainty which surround the therapeutic decision, while variability from purely statistical approaches will mainly depend on the sample size.

Applications

Patients with chronic liver cirrhosis

In an randomized clinical trial studying the efficacy of prednisone (treatment A, n=226) to increase the survival of patients with chronic liver cirrhosis (24), no significant difference was reported compared to patients receiving placebo (treatment B, n=220). Nevertheless, an interaction with the initial prothrombin level, an essential blood clotting glycoprotein, suggested the better efficacy of prednisone in patients with a high prothrombin level (30). Patients with a poorer liver function, i.e. a low prothrombin level, did not benefit from prednisone as their liver were unable to metabolize the high steroid dose, and therefore may be steroid-poisoned. For the purpose of illustration, we aimed to reanalyze the data of this historic clinical trial to define the prothrombin threshold above which patients may receive prednisone, by taking into account that prednisone is a glucocorticosteroid that may be beneficial to liver function (31) whilst causing side effects such as osteopenia, diabetes mellitus, elevation of arterial blood pressure, psychiatric disorders, glaucoma or serious infections (32). We chose the prognostic time τ at 8 years, as the maximal

follow-up time with sufficient at-risk patients for inference. More precisely, 36 prednisone-treated patients and 32 placebo-treated patients were still at risk.

The patient survival probabilities at 8 years were 37% (95% CI from 30% to 45%) for the prednisone-treated patients and 28% (95% CI from 22% to 36%) for the placebo-treated patients. The median prothrombin level was 68% (range from 12% to 134%). The area under the ROC curve for a prognostic up to 8 years (Figure 1) was 0.59 (95% CI from 0.51 to 0.66) reflecting moderate prognostic capacities of prothrombin measurement.

Utilities of the various health states were defined according to the mean Euroqol EQ-5D values published in a systematic review of health-state utilities in liver disease performed by Mc Lernon et al. (33). The EQ-5D is a widely used five-item health-related quality of life descriptive system that provides utility scores for 243 possible health states estimated by applying the time-trade off method over samples of the general population (34,35). The assessed utility score was 0.75 for patients with compensated cirrhosis. We used this value for patients receiving the placebo, i.e. $u_{B,0} = 0.75$. Compared to placebo, we may reasonably assume that the impact of prednisone on the QoL ranges between a 5% decrease ($u_{A,0} = 0.95 \times u_{B,0} = 0.71$) and a 10% decrease ($u_{A,0} = 0.90 \times u_{B,0} = 0.67$). Because the treatment allocation was performed independently of the prothrombin level, the RMSTs used in the equation (3) can be estimated assuming no confounding factors.

The results are presented in Table 1. With a 5% utility decrease due to prednisone, the optimal prothrombin threshold was estimated at 89% (95% CI from 16% to 91%). The corresponding expected utility was 3.54 years in perfect health (QALYs). In comparison, this mean value should have been 3.41 QALYs by treating all the patients with prednisone and 3.32 QALYs by treating all

the patients with placebo. In others words, for 100 patients treated by prednisone if their prothrombin level is higher than 89% and by placebo otherwise, the expected gain was 13 years in perfect health $(100 \times (3.54 - 3.41))$ compared to treating the 100 patients by prednisone. Similarly, the expected gain of the stratified medical decision for a prothrombin at 89% was 22 years in perfect health $(100 \times (3.54 - 3.32))$ compared to treating the 100 patients by placebo. As also indicated in the Table 1, one can note that stratifying the medical decision at 89% results in a comparable RMST than treating systematically by prednisone: 4.80 versus 4.81 years, respectively. Similarly, the difference of RMST when all the patients received the placebo is small: 4.80 versus 4.42 years. That illustrates the importance to not only consider the decision in terms of efficacy, which may not be so pertinent than using QALYs. Twenty-two percent of patients had a prothrombin higher than 89%. In this subgroup and for patients followed up to 8 years, we estimated at 1.7 years the mean increase of time-to-death attributable to prednisone (versus placebo). This was equivalent to a 1 year survival gain in perfect health. In patients with a prothrombin level lower than 89%, the placebo resulted in a similar RMST compared to prednisone. Nevertheless, this equivalent RMST corresponded to an increase of 0.17 QALY due to the side effects of prednisone. One can note that the results were similar with a 10% decrease in the utility due to prednisone treatment. The prothrombin threshold is still estimated at 89% (95% CI from 36% to 99%). Additionally, by exploring others prognostic times τ from 3 years to 8 years, the optimal threshold appeared stable, ranging from 87% to 89%.

From purely statistical approaches, we retained different thresholds (Table 1 and Figure 1). Using the Youden index for instance, the retained threshold of the prognostic marker was 59% (95% CI from 43% to 92%), corresponding to 62% of patients that should be treated by prednisone in a

future stratified medical decision making situation. The corresponding expected utility was 3.42 QALYs for a 5% utility decrease. In other words, for 100 patients treated by prednisone if their prothrombin level is higher than 59% and by placebo otherwise, the expected loss was 12 years in perfect health $(100 \times (3.42 - 3.54))$ compared to the same decision rules with the threshold at 89% we previously reported by using our proposed approach. By a profile likelihood maximization of a Cox model including the treatment, the prothrombin level (1 if the level is higher than κ and 0 otherwise) and the corresponding interaction, we estimated a prothrombin threshold at 54% (95% CI from 36 % to 90%): 72% of patients that should be treated by prednisone. The corresponding expected utility was 3.37 QALYs for a 5% utility decrease, which represents an expected loss for 100 patients equals to 17 years in perfect health $(100 \times (3.37 - 3.54))$ compared to the same decision rules with the threshold at 89%.

Kidney transplant recipients

Obtained from the observational DIVAT cohort, the KTFS (Kidney Transplantation Failure Score) is a score calculated at one year post-transplantation and composed of 8 clinical parameters aiming to predict returns to dialysis within the 8 years post-transplantation (25). The authors proposed a threshold at 4.17 by maximizing both the time-dependent sensitivity and specificity (36), a method close to the Youden index maximization. A randomized clinical trial is currently in progress in order to determine whether the efficiency of the transplantation could be improved by adapting the recipient follow-up, in particular with a higher frequency of visits for patients stratified at high-risk, i.e. with KTFS higher than 4.17 (37).

We reanalyzed the training sample used in the initial paper (25) with the same prognostic time τ at 8 years. In this cohort, all patients received standard care in terms of visit frequency (treatment B). We wondered whether some patients could benefit from a more intensive follow-up (treatment A) regarding their higher risk of return to dialysis, even if one can also expect a decrease of recipients' QoL due to repeated in-hospital visits with negative consequences such as anxiety, depression and distress that have been documented in some diseases (38). This sample was composed of 2169 kidney transplant patients, 182 had returned to dialysis at the end of the follow-up. The graft survival probabilities at 4 and 8 years post-transplantation were respectively 95% (95% CI from 93% to 96%) and 85% (95% CI from 83% to 88%). The median KTFS was equal to 3.73 (range from 1.23 to 15.33). Prognostic capacities at 8 years appeared good with an area under the ROC curve (Figure 2) estimated at 0.78 (95% CI from 0.73 to 0.80).

The mean EQ-5D utility score for a functional kidney transplant was estimated at 0.81 in a meta-analysis by Liem et al. (39). We found only one study reporting the utility after a return in dialysis (40), which was estimated at 65% of the utility of having a functional transplant. More formally, we have: $u_{B,0} = 0.81$ and $u_{A,1} = u_{B,1} = 0.65 \times u_{B,0} = 0.53$. We defined several expected QoL decreases related to a higher frequency of visits compared to standard follow-up: *i)* 10% ($u_{A,0} = 0.90 \times u_{B,0} = 0.73$), *ii)* 5% ($u_{A,0} = 0.95 \times u_{B,0} = 0.77$) and *iii)* 1% ($u_{A,0} = 0.99 \times u_{B,0} = 0.80$). We found no published data supporting the gain in RMST related to the follow-up frequency increase, we therefore hypothesized two possible scenarios: a 5% and a 10% RMST increase compared to standard follow-up. Note that the gain was assumed identical whatever the KTFS threshold.

The results are presented in Table 2. For a 10% decrease in the utility of a transplanted patient, the maximization of the time-dependent expected utility led to maintaining all patients on a standard follow-up regardless the RMST gain ($\kappa^* = 15.33$, the maximum observed value of KTFS). This means that the expected gain of RMST was insufficient to outweigh the consequences due to the increased follow-up frequency in terms of QoL. The corresponding mean RMST and the expected utility were respectively 7.54 years and 6.35 QALYs. Similarly, with a 5% decrease in the utility of a transplanted patient and a 10% RMST increase, the optimal threshold was estimated at 9.34 (95% CI from 7.45 to 15.33) with only 1% of patients having a higher KTFS. Only a very small (but unrealistic) 1% decrease in the utility due to a higher visit frequency could lead to intermediate optimal thresholds. The corresponding mean RMST and the expected utility were respectively 7.95 years and 6.43 QALYs. In addition, we also explored optimal threshold estimation for different prognostic times τ from 2 years to 8 years post-transplantation. Our conclusions were similar: only a 1% utility decrease allows to estimate discriminating thresholds (data not shown). Note that using the Youden index, the threshold was 4.07 (95% CI from 4.05 to 4.43) (Figure 2), close to the threshold 4.17 initially proposed by Foucher et al. (25), leading to 68% of patients with a higher KTFS. But the Youden index does not lead to the threshold that best represent the patients' preferences. For instance, assuming a 10% utility decrease and a 5% RMST increase, for 100 kidney recipients for which a stratified medical decision is made given a KTFS threshold at 4.07, the expected loss was 19 years in perfect health $(100 \times (6.16 - 6.35))$ compared to propose systematically the standard follow-up.

Discussion

The use of prognostic markers for the development of stratified medicine is conditioned upon the determination of clinically relevant thresholds allowing to discriminate patients according to whether they could benefit or not from a treatment. In agreement with recent commentaries on personalized medicine (6), we believe that the clinical utility of such prognostic markers should be evaluated in a patient-centered framework (41), taking into consideration the ultimate objective of biomedical studies: improving patient well-being.

We proposed a threshold estimator by maximizing a time-dependent QALYs-based expected utility function. This approach has several interesting features. Firstly, it takes into account both a marker's prognostic capacities in the presence of censored data and individual preferences over health outcomes. Secondly, our proposal may prove simple and directly applicable to various medical contexts, as illustrated by the applications we provided. It avoids having to directly assess patients' preferences, which is often considered a difficult and costly task (42), since QALYs can be estimated from indirect information about patient preferences, available in the literature for a wide variety of health states. Thirdly, as showed in the applications as well as in the simulation study, the maximization of the proposed time-dependent expected utility function lead to estimating thresholds that are potentially different from those estimated using purely statistical methods, which questions their clinical relevance. Lastly, unlike statistical approaches, our method allows to take into account the consequences of therapeutic uncertainty regarding the possible outcomes of a stratified medicine program.

Subtil et al. (9) recalled that expected utility function for stratified decision making are related to decision curves developed by Vickers et al. (42,43) and initially proposed to compare predictive

models. Decision Curves Analysis (DCA) does not directly compare to our approach since it does not aim at determining an optimal threshold for the marker. DCA does not require explicitly patients' preferences elicitation since it assumes that physicians have knowledge of threshold probability of a disease at which a patient would opt for treatment instead of doing nothing. This threshold probability is associated to the ratio of net benefit of treating a diseased patient to net cost of treating a non-diseased one. In this spirit, Foucher et al. (44) asked physicians for their estimation of the probability threshold for an intensive follow-up program of renal transplant patients. However, physicians may find it difficult to assess this relative weight and their assessments may not be in accordance with patients' preferences (45,46). By contrast, our approach need neither to assume a prior knowledge of patients' preferences (patient's threshold probability) nor to ask physicians about these preferences but uses a well-established and widely used measure of health outcomes.

Besides, the two applications highlight that high prognostic capacities of a marker, summarized for instance through a high area under the ROC curve, is not sufficient to demonstrate its clinical utility for patient-centered stratified medicine. Actually, the KTFS appeared useless at driving the frequency of recipient follow-up, although the corresponding AUC at 0.78 (95% CI from 0.73 to 0.80) can be considered as elevated. Conversely, in the chronic liver cirrhosis application, the ROC curve related to the prothrombin level leads to a moderate AUC at 0.59 (95% CI from 0.51 to 0.66) for a prognostic at 8 years. Nevertheless, a stratified medical program based on the prothrombin threshold may increase patient well-being. Note that the interaction between the marker of interest and the treatment appears as important as the prognostic capacities of the marker and may influenced the threshold estimation.

Our proposal has nevertheless some limitations that are worth mentioning. First, it has been argued that regret minimization offers a more accurate description of physicians' behavior than expected utility (47,48). However, we are concerned here by prescriptive medical decision making and expected utility maximization is still thought of as the best theory to recommend optimal decisions (49). Secondly, implementing our approach requires making assumptions about the potential consequences of stratified medicine. In some ways, this is similar to the exercise made in randomized controlled trial planning to estimate a sample size. Interestingly, Royston et al. recently proposed the use of the difference in RMST to design and analyze clinical trials (50). The anticipation of potential future health outcomes is intrinsic to the exercise of stratified medicine. Thirdly, our time-dependent expected utility (equation (3)) does not discount future QALYs, in contrast to the general formulation (equation (1)). In a normative perspective the choice of the discount rate and the legitimacy of discounting future health effects are controversial (32). Adding a discount factor would be straightforward and would not change the central message of the paper. Finally, when the two treatments are observed, we did not develop the issue related to potential confounders. Based on the recent development we published on adjusted survival curves (51), we are currently developing a confounder-adjusted time-dependent expected utility. To conclude, we have proposed a decision analytic method to determine optimal thresholds of prognostic markers based on the maximization of a time-dependent expected utility function, allowing to overcome the limitation of purely statistical approaches. The package *ROCt* in R has been updated to make this methodology simple for users and available for future applications. Applying such a patient-centered methodology may improve future transfer of novel prognostic scoring systems or markers in clinical practice.

Acknowledgements

We thank the DIVAT scientific council and the members of the clinical research assistant team (M. Kessler, M. Ladrière, JP. Soulillou, C. Legendre, H. Kreis, G. Mourad, V. Garrigue, L. Rostaing, N. Kamar, E. Morelon, F. Buron, S. Le Floch, C. Scellier, V. Eschbach, P. Przednowed, K. Zurbonsen, V. Godel, X. Longy, C. Dagot, F. M'Raiagh) for the collection of the data in the DIVAT cohort.

Funding

This work was supported by the French National Research Agency [ANR-11-JSV1-0008-01, 2011]; and the French National Cancer Institute [INCA 2013-137, 2013].

Declaration of Conflicting Interest

The Authors declare that there is no conflict of interest.

References

1. Baker SG, Sargent DJ. Designing a Randomized Clinical Trial to Evaluate Personalized Medicine: A New Approach Based on Risk Prediction. J Natl Cancer Inst. 2010 Dec 1;102(23):1756–9.

2. Steffen JA, Steffen JS. Driving Forces Behind the Past and Future Emergence of Personalized Medicine. J Pers Med. 2013 Jan 17;3(1):14–22.

3. Ioannidis JPA. Expectations, validity, and reality in omics. J Clin Epidemiol. 2010 Sep;63(9):945–9.

4. Ioannidis JPA. Evolution and Translation of Research Findings: From Bench to Where? PLoS Clin Trials [Internet]. 2006 Nov 17 [cited 2015 Dec 3];1(7). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1851723/

5. Iglesias AI, Mihaescu R, Ioannidis JPA, Khoury MJ, Little J, van Duijn CM, et al. Scientific reporting is suboptimal for aspects that characterize genetic risk prediction studies: a review of published articles based on the Genetic RIsk Prediction Studies statement. J Clin Epidemiol. 2014 mai;67(5):487–99.

6. Joyner MJ, Paneth N. SEven questions for personalized medicine. JAMA. 2015 Sep 8;314(10):999–1000.

7. Youden WJ. Index for rating diagnostic tests. Cancer. 1950 Jan;3(1):32–5.

8. Böhning D, Holling H, Patilea V. A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. Stat Methods Med Res. 2011 Oct 1;20(5):541–50.

9. Subtil F, Rabilloud M. An enhancement of ROC curves made them clinically relevant for diagnostic-test comparison and optimal-threshold determination. J Clin Epidemiol. 2015 Jul;68(7):752–9.

10. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc. 2009 Oct;172(4):729–48.

11. Baker SG, Kramer BS. Peirce, Youden, and Receiver Operating Characteristic Curves. Am Stat. 2007 Nov;61(4):343–6.

12. Pauker SG, Kassirer JP. The Threshold Approach to Clinical Decision Making. N Engl J Med. 1980;302(20):1109–17.

13. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978 Oct;8(4):283–98.

14. Irwin RJ, Irwin TC. A principled approach to setting optimal diagnostic thresholds: where ROC and indifference curves meet. Eur J Intern Med. 2011;22(3):230–4.

15. Cantor SB, Sun CC, Tortolero-Luna G, Richards-Kortum R, Follen M. A comparison of C/B ratios from studies using receiver operating characteristic curve analysis. J Clin Epidemiol. 1999 Sep;52(9):885–92.

16. Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2009 Mar;12 Suppl 1:S5–9.

17. Kind P, Lafata JE, Matuszewski K, Raisch D. The use of QALYs in clinical and patient decision-making: issues and prospects. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2009 Mar;12 Suppl 1:S27–30.

18. Towse A. Net clinical benefit: the art and science of jointly estimating benefits and risks of medical treatment. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2010 Jun;13 Suppl 1:S30–2.

19. Ferguson ND, Scales DC, Pinto R, Wilcox ME, Cook DJ, Guyatt GH, et al. Integrating Mortality and Morbidity Outcomes. Am J Respir Crit Care Med. 2013 Feb 1;187(3):256–61.

20. Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ. 1986 Mar;5(1):1–30.

21. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy Amst Neth. 1990 Dec;16(3):199–208.

22. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002 Mar;21(2):271–92.

23. Ara R, Wailoo A. Using Health State Utility Values in Models Exploring the Cost-Effectiveness of Health Technologies. Value Health. 2012 Sep;15(6):971–4.

24. Schlichting P, Christensen E, Andersen PK, Fauerholdt L, Juhl E, Poulsen H, et al. Prognostic factors in cirrhosis identified by Cox's regression model. Hepatol Baltim Md. 1983 Dec;3(6):889–95.

25. Foucher Y, Daguin P, Akl A, Kessler M, Ladriere M, Legendre C, et al. A clinical scoring system highly predictive of long-term kidney graft survival. Kidney Int. 2010;78(12):1288–94.

26. Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2009 Mar;12 Suppl 1:S5–9.

27. Andersen PK, Perme MP. Pseudo-observations in survival analysis. Stat Methods Med Res. 2010 Feb;19(1):71–99.

28. Efron B. Bootstrap Methods: Another Look at the Jacknife. Ann Stat. 1979;1–26.

29. R Foundation for Statistical Computing. R Development Core Team . R: A language and environment for statistical computing. Vienna, Austria. [Internet]. 2011. Available from: ISBN 3-900051-07-0, URL

30. Christensen E, Schlichting P, Andersen PK, Fauerholdt L, Schou G, Pedersen BV, et al. Updating Prognosis and Therapeutic Effect Evaluation in Cirrhosis with Cox's Multiple Regression Model for Time-Dependent Variables. Scand J Gastroenterol. 1986 Jan 1;21(2):163–74.

31. Mitchison HC, Palmer JM, Bassendine MF, Watson AJ, Record CO, James OF. A controlled trial of prednisolone treatment in primary biliary cirrhosis. Three-year results. J Hepatol. 1992 Jul;15(3):336–44.

32. Swartz SL, Dluhy RG. Corticosteroids: clinical pharmacology and therapeutic use. Drugs. 1978 Sep;16(3):238–55.

33. McLernon DJ, Dillon J, Donnan PT. Health-state utilities in liver disease: a systematic review. Med Decis Mak Int J Soc Med Decis Mak. 2008 Aug;28(4):582–92.

34. Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. Health Econ. 1996 Apr;5(2):141–54.

35. Chevalier J, de Pouvourville G. Valuing EQ-5D using time trade-off in France. Eur J Health Econ HEPAC Health Econ Prev Care. 2013 Feb;14(1):57–66.

36. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337–44.

37. Foucher Y, Meurette A, Daguin P, Bonnaud-Antignac A, Hardouin J-B, Chailan S, et al. A personalized follow-up of kidney transplant recipients using video conferencing based on a 1-year scoring system predictive of long term graft failure (TELEGRAFT study): protocol for a randomized controlled trial. BMC Nephrol. 2015;16:6.

38. Kew FM, Galaal K, Manderville H. Patients' views of follow-up after treatment for gynaecological cancer. J Obstet Gynaecol J Inst Obstet Gynaecol. 2009 Feb;29(2):135–42.

39. Liem YS, Bosch JL, Hunink MGM. Preference-based quality of life of patients on renal replacement therapy: a systematic review and meta-analysis. Value Health J Int Soc Pharmacoeconomics Outcomes Res. 2008 Aug;11(4):733–41.

40. Girardi V, Schaedeli F, Marti H-P, Frey FJ, Uehlinger DE. The willingness of patients to accept an additional mortality risk in order to improve renal graft survival. Kidney Int. 2004 Jul;66(1):375–82.

41. Walter SD, Turner R, Macaskill P, McCaffery KJ, Irwig L. Beyond the treatment effect: Evaluating the effects of patient preferences in randomised trials. Stat Methods Med Res. 2014 Sep 11;0962280214550516.

42. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak Int J Soc Med Decis Mak. 2006 Dec;26(6):565–74.

43. Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. Med Decis Mak Int J Soc Med Decis Mak. 2008 Feb;28(1):146–9.

44. Foucher Y, Giral M, Soulillou JP, Daures JP. Cut-off estimation and medical decision making based on a continuous prognostic factor: the prediction of kidney graft failure. Int J Biostat [Internet]. 2012;8(1). Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=22499724

45. Sonnenberg A. Patient–physician discordance about benefits and risks in gastroenterology decision-making. Aliment Pharmacol Ther. 2004 Jun 1;19(12):1247–53.

46. Yuan Z, Levitan B, Burton P, Poulos C, Hauber AB, Berlin JA. Relative importance of benefits and risks associated with antithrombotic therapies for acute coronary syndrome: patient and physician perspectives. Curr Med Res Opin. 2014 Sep 1;30(9):1733–41.

47. Djulbegovic B, Hozo I, Schwartz A, McMasters KM. Acceptable regret in medical decision making. Med Hypotheses. 1999 Sep;53(3):253–9.

48. Tsalatsanis A, Barnes LE, Hozo I, Djulbegovic B. Extensions to regret-based decision curve analysis: an application to hospice referral for terminal patients. BMC Med Inform Decis Mak. 2011;11.

49. Wakker PP. Lessons learned by (from?) an economist working in medical decision making. Med Decis Mak Int J Soc Med Decis Mak. 2008 Oct;28(5):690–8.

50. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013;13(1):152.

51. Le Borgne F, Giraudeau B, Querard AH, Giral M, Foucher Y. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. Stat Med. 2015 Oct 29;

Figure 1: Receiver-Operating Characteristic (ROC) curves for 8-year predictions to evaluate the prognostic capacity of the prothrombin marker among patients with chronic liver cirrhosis (n=446).

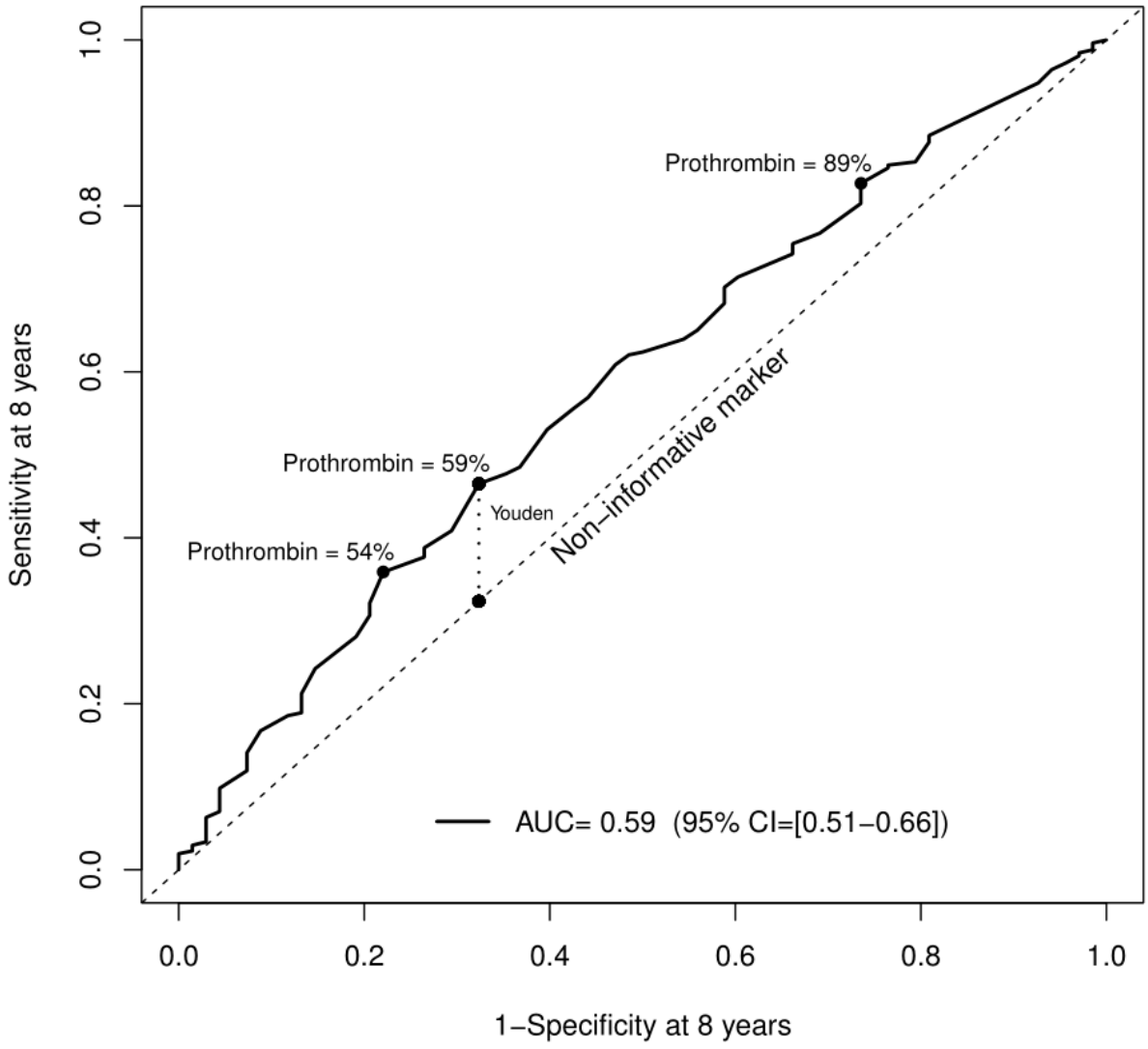

Figure 2: Receiver-Operating Characteristic (ROC) curves for 8-year predictions to evaluate the prognostic capacity of the KTFS among kidney transplant recipients (n=2169).

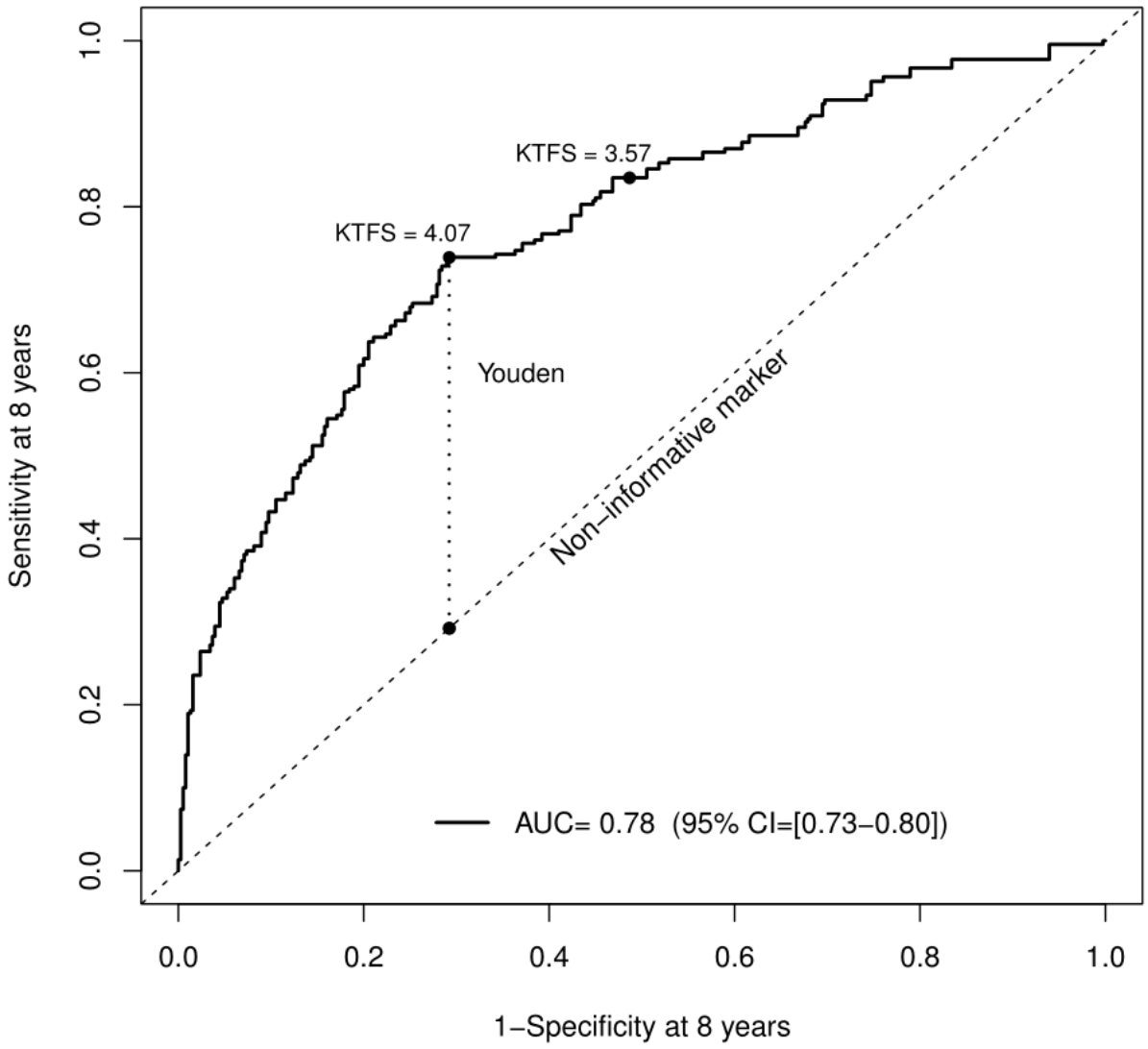

Table 1: Results of the application on patients suffering chronic liver cirrhosis for scenarios with 5% and 10% utility decrease due to prednisone treatment compared to placebo.

Utility decrease	Threshold κ^*	Percentage with $X > \kappa^*$	RMST change (years) over misclassified treated patients		QALY change (years) over misclassified treated patients		Mean RMST (in years)	Expected utility (in QALYs)
			$X > \kappa^*$	$X \leq \kappa^*$	$X > \kappa^*$	$X \leq \kappa^*$		
5%	Maximum: 134	0%	-	-	-	-	4.42	3.32
	QALYs-based expected utility : 89 [16-91][1]	22%	1.67	-0.01	0.99	0.17	4.80	3.54
	Youden index : 59 [43-92][1]	62%	0.53	-0.23	0.18	-002	4.74	3.42
	Profile likelihood maximisation : 54 [36-92][1]	72%	0.36	-0.47	0.06	-0.21	4.69	3.37
	Minimum : 12	100%	-	-	-	-	4.81	3.41
10%	Maximum: 134	0%	-	-	-	-	4.42	3.32
	QALYs-based expected utility : 89 [13-99][1]	22%	1.67	-0.01	0.74	0.34	4.80	3.48
	Youden index : 59 [43-92][1]	62%	0.53	-0.23	-0.04	0.13	4.74	3.29
	Profile likelihood maximisation : 54 [36-92][1]	72%	0.36	-0.47	-0.16	-0.07	4.69	3.21
	Minimum : 12	100%	-	-	-	-	4.81	3.22

[1] CI95% calculated from 2000 bootstrap samples

Table 2: Results of the application on kidney transplant recipients for scenarios with 1%, 5% and 10% utility decrease due to prednisone treatment associated with 5% or 10% RMST increase due to a higher consultation frequency.

Utility decrease	RMST increase	Threshold κ^* [Bootstrap 95% CI][1]	Percentage with $X > \kappa^*$	RMST change (years) over observed patients $X > \kappa^*$	QALY change (years) over observed patients $X \leq \kappa^*$	Mean RMST (in years)	Expected utility (in QALYs)
10%	10%	15.33 [10.85-15.33]	0%	-	-	7.54	6.35
	5%	15.33 [11.59-15.33]	0%	-	-	7.54	6.35
5%	10%	9.34 [7.45-15.33]	1%	0.39	-0.06	7.54	6.35
	5%	10.31 [8.36-15.33]	1%	0.18	-0.10	7.56	6.35
1%	10%	3.57 [2.96-3.80]	57%	0.73	0.14	7.95	6.43
	5%	1.23 [1.23-2.03]	100%	0.38	0.04	7.91	6.39
10%	10%	Youden index : 4.07 [4.05-4.43]	68%	0.70	-0.43	7.80	6.19
	5%			0.35	-0.50	7.67	6.16
5%	10%			0.70	-0.12	7.80	6.30
	5%			0.35	-0.20	7.67	6.27
1%	10%			0.70	0.13	7.80	6.40
	5%			0.35	0.04	7.67	6.36

[1] CI95% calculated from 2000 bootstrap samples

Table 3: Results of the simulation study performed on 1000 samples of 1000 patients.

			$u_A = u_B = 0.7$		$u_A = 0.6$ $u_B = 0.7$		$u_A = 0.5$ $u_B = 0.7$	
Interaction between marker and treatment β_{XZ}	Treatment effect β_Z	Estimator	Mean of estimated κ	SE	Mean of estimated κ	SE	Mean of estimated κ	SE
0	0	QALYs-based expected utility	144.10	75.56	282.13	34.07	304.15	27.31
		Youden index	141.67	25.06	141.67	25.06	141.67	25.06
		Profile likelihood maximization	137.35	36.10	137.35	36.10	137.35	36.10
	-0.5	QALYs-based expected utility	36.71	18.10	243.88	45.73	297.61	29.92
		Youden index	142.39	25.99	142.39	25.99	142.39	25.99
		Profile likelihood maximization	137.78	36.79	137.78	36.79	137.78	36.79
	-1.25	QALYs-based expected utility	25.92	7.76	91.61	45.58	281.49	34.32
		Youden index	142.91	28.06	142.91	28.06	142.91	28.06
		Profile likelihood maximization	141.88	38.97	141.88	38.97	141.88	38.97
-0.004	-0.5	QALYs-based expected utility	31.49	12.51	126.48	39.31	276.78	34.54
		Youden index	141.11	33.98	141.11	33.98	141.11	33.98
		Profile likelihood maximization	145.61	43.40	145.61	43.40	145.61	43.40
	-1.25	QALYs-based expected utility	25.36	7.11	62.67	30.32	264.40	36.90
		Youden index	142.69	33.30	142.69	33.30	142.69	33.30
		Profile likelihood maximization	144.19	45.17	144.19	45.17	144.19	45.17
-0.008	-0.5	QALYs-based expected utility	29.18	10.19	85.57	31.53	260.56	36.86
		Youden index	141.04	34.53	141.04	34.53	141.04	34.53
		Profile likelihood maximization	141.12	38.93	141.12	38.93	141.12	38.93
	-1.25	QALYs-based expected utility	24.89	6.37	52.59	24.44	252.98	38.37
		Youden index	144.02	34.31	144.02	34.31	144.02	34.31
		Profile likelihood maximization	140.93	43.13	140.93	43.13	140.93	43.13

Appendix

The expected utility $\psi_\tau(\kappa)$, as defined by the equation (1), may be developed as follows by considering the number of QALYs (equation (2)):

$$\psi_\tau(\kappa) = \sum_{g\in\{X>\kappa, X\le\kappa\}} \sum_{j\in\{0,1\}} P(g, D(\tau)=j) Q(\tau|g, D(\tau)=j)$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \{P(g, D(\tau)=0) Q(\tau|g, D(\tau)=0) + P(g, D(\tau)=1) Q(\tau|g, D(\tau)=1)\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \Bigg\{ P(g, D(\tau)$$

$$= 0) \left[\int_0^\tau u(t|g, D(t)=0) P(D(t)=0|g, D(\tau)=0) \right.$$

$$\left. + u(t|g, D(t)=1) P(D(t)=1|g, D(\tau)=0) dt \right]$$

$$+ P(g, D(\tau)=1) \left[\int_0^\tau u(t|g, D(t)=0) P(D(t)=0|g, D(\tau)=1) \right.$$

$$\left. \left. + u(t|g, D(t)=1) P(D(t)=1|g, D(\tau)=1) dt \right] \right\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \Bigg\{ P(g, D(\tau)=0) \left[\int_0^\tau u(t|g, D(t)=0) dt \right]$$

$$+ P(g, D(\tau)=1) \left[\int_0^\tau u(t|g, D(t)=0) P(D(t)=0|g, D(\tau)=1) \right.$$

$$\left. \left. + u(t|g, D(t)=1) P(D(t)=1|g, D(\tau)=1) dt \right] \right\}$$

We assumed a constant utility over time, i.e. $u(t|g, D(t) = l) = u_{g,l}$ for all t. Thus, the expected utility $\psi_\tau(\kappa)$ can be written as:

$$\psi_\tau(\kappa) = \sum_{g\in\{X>\kappa, X\le\kappa\}} \Bigg\{ P(g, D(\tau) = 0) \left[\int_0^\tau u_{g,0} dt \right] + P(g, D(\tau) = 1) \left[\int_0^\tau u_{g,0} P(D(t) = 0|g, D(\tau) = 1) + u_{g,1} P(D(t) = 1|g, D(\tau) = 1) dt \right] \Bigg\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \Bigg\{ \int_0^\tau u_{g,0} \big(P(g, D(\tau) = 0) + P(g, D(\tau) = 1) P(D(t) = 0|g, D(\tau) = 1) \big) dt + \int_0^\tau u_{g,1} P(g, D(\tau) = 1) P(D(t) = 1|g, D(\tau) = 1) dt \Bigg\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \Bigg\{ u_{g,0} \int_0^\tau \big(P(g, D(\tau) = 0) + P(D(t) = 0, g, D(\tau) = 1) \big) dt + u_{g,1} \int_0^\tau P(D(t) = 1, g, D(\tau) = 1) dt \Bigg\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ u_{g,0} \int_0^\tau P(D(t) = 0, g) dt + u_{g,1} \int_0^\tau P(D(t) = 1, g) dt \right\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g) \times \left[u_{g,0} \times \int_0^\tau P(D(t) = 0|g)\, dt + u_{g,1} \times \int_0^\tau P(D(t) = 1|g)\, dt \right] \right\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g) \times \left[u_{g,0} \times \int_0^\tau S(t|g)\, dt + u_{g,1} \times \int_0^\tau 1 - S(t|g)\, dt \right] \right\}$$

$$= \sum_{g\in\{X>\kappa, X\le\kappa\}} \left\{ P(g) \times \left[u_{g,0} \times E(\min(T, \tau)\,|g) + u_{g,1} \times \big(\tau - E(\min(T, \tau)\,|g) \big) \right] \right\}$$