BCPT Basic & Clinical Pharmacology & Toxicology

Basic & Clinical Pharmacology & Toxicology Doi: 10.1111/j.1742-7843.2012.00885.x

Data-Driven Assessment of the Association of Polymorphisms in 5-Fluorouracil Metabolism Genes with Outcome in Adjuvant Treatment of Colorectal Cancer

Sinan B. Sarac[1,2], Christian H. Rasmussen[1,2], Shoaib Afzal[3], Steffen Thirstrup[4], Søren A. Jensen[5], Morten Colding-Jørgensen[1], Henrik E. Poulsen[6,7,8] and Erik Mosekilde[2]

[1]Novo Nordisk A/S, Bagsværd, Denmark, [2]Department of Physics, Technical University of Denmark, Lyngby, Denmark, [3]Department of Clinical Biochemistry, Herlev Hospital, Herlev, Denmark, [4]Danish Medicines Agency, Copenhagen, Denmark, [5]Department of Oncology, Herlev Hospital, Herlev, Denmark, [6]Laboratory of Clinical Pharmacology, Rigshospitalet, Copenhagen, Denmark, [7]Department of Clinical Pharmacology, Bispebjerg Hospital, Copenhagen, Denmark, and [8]Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

(Received 19 December 2011; Accepted 19 March 2012)

Abstract: A major challenge in the assessment of medicines, treatment options, etc., is to establish a framework for the comparison of risks and benefits of many different types and magnitudes, a framework that at the same time allows a clear distinction between the roles played by the statistical analyses of data and by judgements based on personal experience and expertise. The purpose of this study was to demonstrate how clinical data can be weighted, scored and presented by the use of an eight-step data-driven benefit–risk assessment method, where two genetic profiles are compared. Our aim was to present a comprehensive approach that is simple to apply, allows direct comparison of different types of risks and benefits, quantifies the clinical relevance of data and is tailored for the comparison of different options. We analysed a cohort of 302 patients with colorectal cancer treated with 5-Fluorouracil (5-FU). Endpoints were cure rate, survival rate, time-to-death (TTD), time-to-relapse (TTR) and main adverse drug reactions. Multifactor dimensionality reduction (MDR) was used to identify genetic interaction profiles associated with outcome. We have been able to demonstrate that a specific MDR-derived combination (the MDR-1 group) of dihydropyrimidine dehydrogenase and thymidylate synthase polymorphisms is associated with increased and clinically significant difference for cure and survival rates, TTD and probably also for TTR, which are seen as the most important endpoints. An inferior profile was observed for severe myocardial ischaemia. A probably inferior profile was seen for severe arthralgia/myalgia and severe infections. A clear superior profile was seen for severe mucositis/stomatitis. The proposed approach offers comprehensive, data-driven assessment that can facilitate decision processes, for example, in a clinical setting. It employs descriptive statistical methods to highlight the clinically relevant differences between options.

Significant resources are invested by regulatory agencies, pharmaceutical companies, hospitals and other health-related organisations to analyse the benefit–risk relationships of different drugs, treatments, procedures, etc. It appears, however, that the current approach for such analyses is inadequate in several respects and that the outcome often depends in a somewhat opaque manner on the experience and expertise of the persons performing the analysis. In recognition of these problems, there is presently a strong and continuously growing activity in the field of 'biomedical benefit–risk analysis' with the aim of developing more structured, transparent and comprehensive methods [1–10]. Many of these efforts are inspired by multi criteria decision analysis (MCDA) [11], but aim to place the analysis in a broader context.

It is important that information about the selection of evaluation criteria, about the magnitude, relative frequency, mutual dependences, etc., of the various benefits and risks, about unexplained observations, etc. is retained to the final analysis. However, benefits and risks are recorded differently in trials performed by different methods and different statistical tools, and the evidence in support of the different observations may vary significantly. At crucial stages of the analysis, decisions or interpretations may therefore depend on input from key individuals. In this way, the final assessment becomes subjective and it may differ, for instance, between regulatory agencies and patient organisations.

While statistical significance of the obtained data is a necessary criterion for a biomedical decision, it is not sufficient. The clinical relevance of the observed effects, that is, the magnitude of the effect observed in individual patients and the fraction of the patient population in which particular benefits and risks occur must be characterised as well [12–14]. It seems to us that the focus is currently on statistical significance rather than on clinical relevance of the data. This is not justified, and both aspects must be considered in a well-structured benefit–risk analysis.

In response to these problems, we have developed a transparent and flexible method tailored for the comparison of two different options (i.e. treatments, drugs, genetic polymorphisms) [15]. The main ideas of this method are (i) to bring the various forms of information onto a common scale that allows direct comparison between, for instance, a commonly

Author for correspondence: Sinan B. Sarac, Novo Nordisk A/S, Novo Allé, DK-2880 Bagsværd, Denmark (fax 44889195, e-mail sbsarac@gmail.com).

occurring but small benefit and a rare, but significant risk, (ii) to retain all essential pieces of information to the final decision, (iii) to present this information in an intuitive diagrammatic form that can represent different dimensions of the available information, for example, type, occurrence and clinical relevance, and (iv) to highlight differences between different stakeholders with regard to their view on weighting and scoring of the various criteria. Different assessments can be made for different stakeholders and can later be compared on point-to-point basis, so that differences in opinion are clearly seen.

This manuscript builds on two recent papers by the present authors: (i) Afzal *et al.* [16] and (ii) Sarac *et al.* [15]. The purpose of the study is to demonstrate the method described by Sarac *et al.* [15] by analysing the results obtained by Afzal *et al.* [16] in a study of colorectal cancer treatment. We illustrate how these data can be weighted, scored and communicated in a structured and transparent manner with focus on the quantification of the clinical relevance and on general tendencies in the data.

The work of Afzal *et al.* focuses on 5-Fluorouracil (5-FU), which is widely used to treat solid tumours including colorectal cancers. The cytotoxicity of 5-FU depends primarily on two active metabolites: (i) Fluorodeoxyuridine monophosphate (5-FdUMP) inhibits the thymidylate synthase (TYMS) enzyme [17], (ii) Fluorouridine triphosphate (5-FUTP) impairs RNA function and thereby induces cell toxicity [18,19]. The inhibition of the TYMS enzyme is dependent on and enhanced by intracellular 5,10-methylenetetrahydrofolate [17,20].

Increased sensitivity of cancer cell lines to 5-FU is correlated with decreased expression or activity of dihydropyrimidine dehydrogenase (DPYD), methylenetetrahydrofolate reductase (MTHFR) and TYMS, and increased activity or expression of orotate phosphoribosyltransferase or uridine monophosphate synthetase (OPRT or UMPS) [21–26]. Studies investigating the association of TYMS, DPYD, OPRT and MTHFR polymorphisms or expression with survival in adjuvant 5-FU-based treatment of colorectal cancer have yielded contradictory results, especially regarding TYMS and MTHFR [27–40].

Most studies investigate the association of individual polymorphisms with disease-free survival (DFS) or overall survival (OS) [37]. Theoretically, the 5-FU metabolic phenotype is better explained by multi-gene and pathway-oriented analysis rather than single-gene analysis [41]. A specific multifactor dimensionality reduction (MDR)-derived combination of DPYD and TYMS, variable number of tandem repeats (VNTR) polymorphisms was observed to be associated with increased DFS [16].

Materials and Methods

The data and methods (MDR) described are those of a recent study by Afzal *et al.* [16]. The study by Afzal *et al.* analysed data from 302 patients, where 10 polymorphisms in genes involved in 5-FU pharmacodynamics and pharmacokinetics had been studied. MDR, a nonparametric method, was used to identify genetic interaction profiles associated with the outcome [16]. A MDR analysis of all the polymorphisms and the functional classifications of MTHFR, TYMS and DPYD, have previously shown that variant alleles in DPYD and the TYMS VNTR polymorphism are associated with improved DFS [16].

The median follow-up was 5 years (up to 11 years). Last follow-up date was 30 August 2007. All patients were Caucasian with Dukes' stage B2 and C treated at Rigshospitalet (Copenhagen University Hospital) with surgery and the Mayo regimen (Levofolinate 10 mg/m^2, 5-FU 425 mg/m^2/day, days 1–5, every 28 days, six cycles). DNA was isolated from formalin-fixed paraffin-embedded tumour tissue, with maximum 50% normal tissue. Clinical data and tumour pathology were reviewed retrospectively [16].

Table 1 shows that 302 patients who were eligible for our study distributed according to their MDR classification. The MDR-1 group consists of patients with the combination of variant alleles in the DPYD gene and the TYMS VNTR polymorphism, selected by the MDR algorithm as being associated with improved DFS [16]. For 33 patients, the quality of DNA obtained from the tissue was so poor that the polymorphisms could not be detected.

The data were assessed by the data-driven benefit–risk assessment method as described by Sarac *et al.* [15]. Figure 1 shows the eight successive steps comprising the overall framework. The proposed methodology is a structured approach to the assessment of clinical data that can be briefly characterised by first defining the decision context. Hereafter, the clinical relevance is defined and quantified, and criteria are identified, weighted and scored. Finally, the results are visualised and communicated in tornado-like diagrams and an overall conclusion is drawn. The proposed methodology provides a bolstering structure around the statistical data analysis.

Let us outline the different steps in a little more detail:

Step 1 – Decision context. Every aspect of the decision context must be explicitly described. This involves several issues, for example, the

Table 1.
Clinical data on the examined cohort. Based on the results from [16].

	N = 302
Age at diagnosis (median years, range)	61 (19–85)
Sex	
Male	151 (50%)
Female	151 (50%)
Median follow-up (years)	5.3 (0.1–11.3)
DFS	
Events	142 (47%)
Censored	160 (53%)
OS	
Events	127 (42%)
Censored	175 (58%)
Stage	
B	37 (12%)
C	265 (88%)
Tumour grade	
1	91 (30%)
2	129 (43%)
3	82 (26%)
Missing	2 (1%)
Tumour site	
Colon	246 (81%)
Rectum	56 (19%)
MDR-1	111
MDR-0	158
Missing	33[1]

DFS, disease-free survival; MDR, multifactor dimensionality reduction; OS, overall survival.
[1]DNA material was of so poor quality that the polymorphism could not be detected.

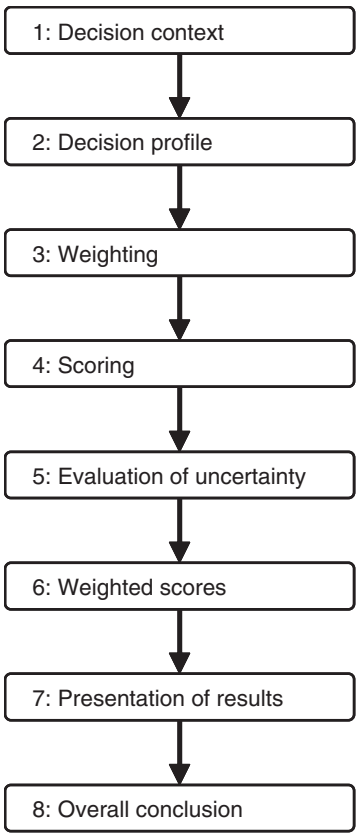

Fig. 1. The eight successive steps in the described benefit–risk assessment method.

stakeholders to be involved in the assessment are identified, aims and goals are defined, and all relevant information to support the assessment is identified [8,42]. We propose that clinical relevance is based on the proportion of patients who experience a specified treatment effect and that clinical relevance is defined and justified from assessment to assessment based on the decision context. This context, therefore, needs to be as clear as possible leaving little room for speculations or questions.

Step 2 – Decision profile. The criteria to be used in the assessment are defined. A criterion is a specific parameter that describes a measurable variable. The selected criteria are common for all groups in the assessment. Each criterion should be based on a measured parameter in a clinical data set. Loss of information must be minimized [3]. The specific selection of criteria to be included in the assessment must be justified for documentation and communication purposes.

Step 3 – Weighting. Criteria are weighted on a value scale to enable comparison in step 3. The weights are based on the relative importance of a difference between two options. Each criterion is assigned a weight/importance of 1 (low, or less important), 2 (medium, or important) or 3 (high, or very important). The weights are comparable across all criteria in the assessment, meaning that two criteria with the same weight are equally important. Weighting is independent of the data sets. All weights must also be justified for documentation and communication purposes. The justification aspect creates the basis for a transparent and credible assessment. Different stakeholders (e.g. clinicians, patients, pharmaceutical companies, health authorities, etc.) may have different opinions about the weights.

Step 4 – Scoring. In this step, the performance of the two groups for each of the selected criteria is assessed. A numerical value is assigned for each criterion, and MDR-1 is scored relative to the comparator, MDR-0, on a simple and transparent scale: −1 (inferior), 0 (non-inferior or equivalent) and +1 (superior). Different types of data require different types of methods for scoring. For continuous variables, such as time-to-relapse (TTR) and time-to-death (TTD), the scoring method is based on difference distributions [43]. A clinically relevant difference is defined as a specified proportion of the patients in a trial experiencing the effects of a treatment relative to a comparator. In this context, clinical relevance is defined as 12 of 20 (=60%) patients in one of the two groups showing a difference for continuous data, for example, time-to-death or time-to-relapse, in order for the difference to be considered clinically significant and relevant. The choice of the ratio is based on the severity of the disease, where even a small difference can have a clinical significance.

In the assessment of event criteria (e.g. responder rate, rare adverse events, etc.), the scoring method is based on confidence intervals (CI). Some events occur only once per patient (e.g. withdrawal, death, etc.). To simplify, we may assume that such events occur independently, with the same probability for each patient. The probability is assumed to depend only on the intervention and in this case, *genetic polymorphisms.* The number of events in a study is then assumed to be binomially distributed. We use the confidence intervals, calculated by the exact method of Clopper and Pearson [44], for scoring. The parameter of the binomial distribution, that is, the probability of a single event for one patient, is calculated based on data. This parameter is a continuous variable for which the scoring is decided by performing two-one-sided binomial tests at, for example, a 2/3 confidence level. This is equivalent to combining two-one-sided 66.7% confidence intervals into a scoring interval and checking for overlap of the intervals. If such an overlap is found, the drug and comparator are deemed non-inferior. This is not an estimation of the uncertainty on the score, but rather on the binomial parameter, just like a mean and standard deviation for a continuous variable.

However, for cure and survival rate, the confidence level is lowered to 60% in accordance with the argumentation used for TTD and TTR. A low confidence level is chosen to capture tendencies in data that otherwise would not have been captured with the conventional level of confidence of 95%. We are aware that the lower the confidence level is, the higher the risk of a type 1 error will be (i.e. the risk that a null hypothesis is incorrectly rejected, when it is actually true).

Step 5 – Evaluation of uncertainty. Scores must be reviewed critically [3,45]. A systematic approach to uncertainty is pivotal and several aspects, for example, methodological flaws, negative studies, etc., can be evaluated subjectively, as part of a qualitative evaluation. The evaluation of uncertainty is integrated in the data-driven scoring method in a simple way. In case of any uncertainty, the score may be given as an interval. The interval −1 to +1 should be chosen, when the available information can only be regarded as defining trends, and re-sampling cannot be justified because of sparse data. Any assignment of interval scores must be justified.

Step 6 – Weighted scores. Weights and scores are multiplied to create a single value for each criterion that captures both its importance and performance. The weighted scores can be used to create overall benefit–risk assessment in the case of multiple trials.

Step 7 – Visualisation. The results are visualised in standardised diagrams. The impact of correlations is limited by avoiding addition and/or multiplication of criteria. Differences between two options are preserved throughout the assessment, and uncertainty and evidence of data are incorporated in both a qualitative and quantitative manner.

Step 8 – Overall conclusion. An overall conclusion is performed as the last step in the assessment. However, all the important discussions related to the importance of criteria, uncertainty of the data, etc., are conducted prior to the creation of tornado-like diagrams. The tornado-like diagrams convey all these small decisions and discussions, and hereby facilitate the overall decision-making process.

Results

Decision context.

The actual assessment is performed from a clinical point of view. Based on the MDR classification, the aim is to investigate the difference in the association of polymorphisms in 5-Fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. We will perform comparison between the two MDR groups on the basis of cure rate, survival rate, TTD, TTR and main adverse drug reactions.

Decision profile.

Based on the decision context, the following criteria are chosen: Cure rate, survival rate, TTD, TTR, infections, bleedings, mucositis/stomatitis, nausea/vomiting, hand-foot-skin syndrome, diarrhoea, arthralgia/myalgia, myocardial ischaemia and fatigue, as seen in table 2. The selected criteria are considered to represent the most relevant aspects in the evaluation of differences between options in the treatment of colorectal cancer.

Weighting.

As with the selection of criteria, all weights must be justified for documentation and communication purposes. In this study, as mentioned previously, weighting is performed from a clinical point of view. Cure rate, survival rate, TTR and TTD are primary endpoints, because both patients and clinicians are interested in overall survival or cure. These criteria are therefore of high importance (weight 3), meaning that a difference between the two MDR groups will have major clinical implications.

Table 2.
The selected criteria and their weights. High importance = 3, medium importance = 2 and low importance = 1.

Criterion	Weight
Cure rate	3
Survival rate	3
TTD	3
TTR	3
Infection	2
Myocardial ischaemia	2
Bleeding	2
Mucositis/stomatitis	2
Hand-foot skin syndrome	2
Diarrhoea	2
Arthralgia/myalgia	1
Fatigue	1
Nausea/vomiting	1

TTD, time-to-death; TTR, time-to-relapse.

The criteria infections, such as myocardial ischaemia, bleedings, mucositis/stomatitis, hand-foot skin syndrome and diarrhoea, are all considered medium important (weight 2), meaning that a difference between two options will probably have clinical implications, for example, if the performance for the high-importance criteria is equal between the two options. The medium importance criteria are often difficult to treat and can have consequences for the patients, occasionally very serious to fatal.

The low importance criteria (weight 1), arthralgia/myalgia, fatigue and nausea/vomiting, can often be treated, for example, pharmacologically. Although, these events can be very severe, they are rarely life-threatening.

Scoring.

Different types of data may require different methods for scoring. For continuous data such as TTR and TTD, distributions are created for the MDR-0 and MDR-1 groups, as seen in fig. 2. There is no follow-up after 8 years, where a considerable number of patients still are alive. We need to know the entire distribution to be able to calculate the difference distribution between two distributions [43]. The Kaplan–Meier curves are therefore fitted to Weibull distributions that are often used to describe and analyse survival data. This distribution seems to also offer a reasonable description of our data. There are other methods for the analysis of survival functions, for example, Mantel–Haenszel methods and hazard ratios, and the proposed methodology is to be seen as an example of these.

Figure 2A shows the TTD as a function of MDR-1 (red) and MDR-0 (blue), respectively. As seen in fig. 2A, a substantial number of patients survive more than 8 years in each group, and the dashed lines represent the 'survival rate' of colorectal cancer for the two MDR classifications, where 67/111 (60%) patients in the MDR-1 group survive compared with 52/158 (33%) patients in the MDR-0 group. The curves are normalised (data not shown), to represent time-to-death for all patients dying during the study. The difference distribution in fig. 2B shows that the patients dying during the study period, 64% in the MDR-1 group have a longer time-to-death relative to the patients dying in the MDR-0 group. Figure 2C,D shows the TTR and is to be interpreted in the same manner. The dashed line represents the 'cure rate', where 66/111 (59%) patients in the MDR-1 group are cured compared to 62/158 (39%) patients in the MDR-0 group. Figure 2D shows that 55% of the patients surviving in the MDR-1 group have longer time-to-relapse relative to the patients surviving in the MDR-0 group.

The dashed lines in fig. 2A,C represent and estimate the levels of survival and cure rates. In table 3, the number of

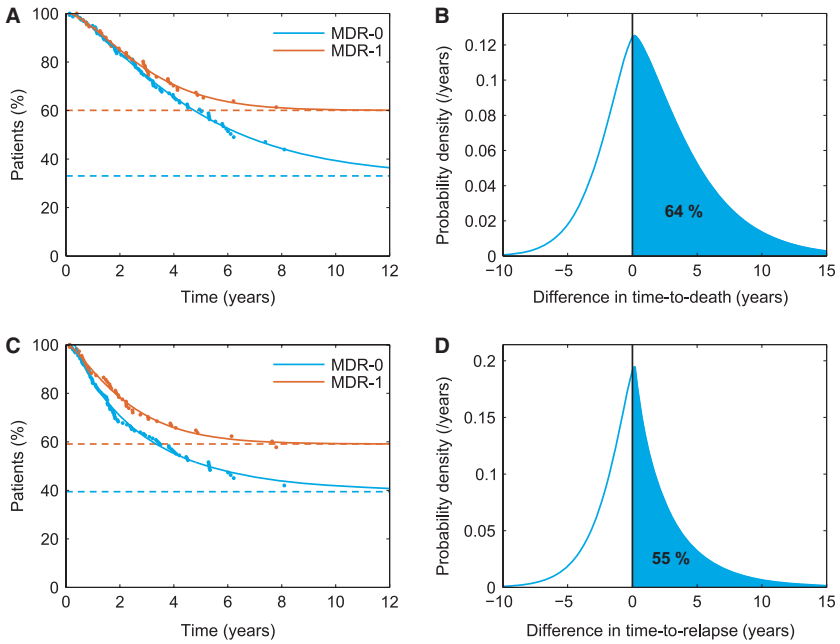

Fig. 2. Data-driven scoring based on difference distributions. (A) Time-to-death (TTD) curves for the MDR-0 and MDR-1 groups, (B) difference distribution for TTD, (C) Time-to-relapse (TTR) curves, (D) difference distribution for TTR. The dashed lines represent the levels of survival and cure rates. 67/111 (60%) patients in the MDR-1 group survive compared to 52/158 (33%) patients in the MDR-0 group, while 66/111 (59%) patients in the MDR-1 group are cured compared to 62/158 (39%) patients in the MDR-0 group.

Table 3.
Number of grade 3–4 toxicity of adjuvant chemotherapy based on the MDR classification, and number of cured and survivors.

Criterion	MDR-1 (N = 111)	MDR-0 (N = 158)
Cure rate	66	62
Survival rate	67	52
Infection	7	6
Myocardial ischaemia	3	0
Bleedings	0	0
Mucositis/stomatitis	10	25
Hand-foot skin syndrome	2	3
Diarrhoea	17	27
Arthralgia/myalgia	2	1
Fatigue	0	4
Nausea/vomiting	5	8

MDR, multifactor dimensionality reduction.

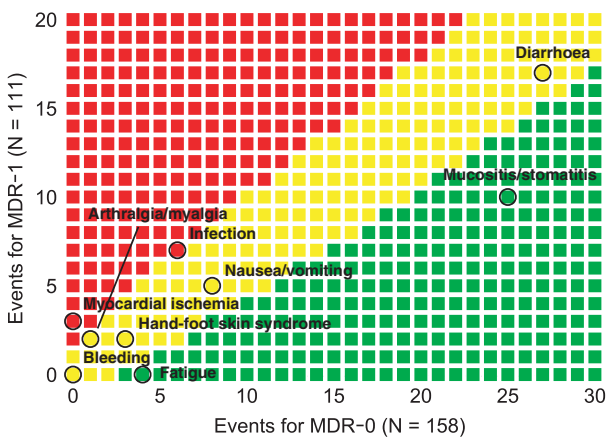

Fig. 3. Data-driven scoring based on confidence interval scoring. Scoring table for the grade 3–4 adverse drug reaction cases. The colour red indicates that MDR-1 is inferior. Yellow represents non-inferiority and green superiority. MDR-1 shows a superior profile with regard to mucositis/stomatitis and fatigue, while inferior for myocardial ischaemia and infections. For all other criteria, MDR-1 is non-inferior.

patients that will either survive the disease or be cured from it is shown.

Toxicity was recorded and graded according the Common Toxicity Criteria (CTC) (National Cancer Institute, Common Toxicity Criteria version 2.0) [16]. We proceed to dichotomously classify toxicity into none-to-moderate (grade 0–2) and severe (grade 3–4) toxicity for the MDR classification. We have chosen to focus on differences in the severity of adverse drug reactions between the two MDR classifications.

Data from table 3 are visualised in the scoring table shown in figs 3 and 4. The green area represents the values at which the confidence intervals based on the Clopper–Pearson method do not overlap with 66.7% confidence limit (60% confidence limit for cure and survival rate), and where the probability of an event hence is lower for MDR-1. The red area represents the values at which the confidence intervals also do not overlap, but where the probability of an event is higher for MDR-1. The yellow area represents the values at which the confidence intervals do overlap, and where one cannot conclude whether or not there is a difference in the probability of an event between the two MDR groups.

Figure 3 shows that patients in the MDR-1 group have a superior profile with regard to mucositis/stomatitis and fatigue.

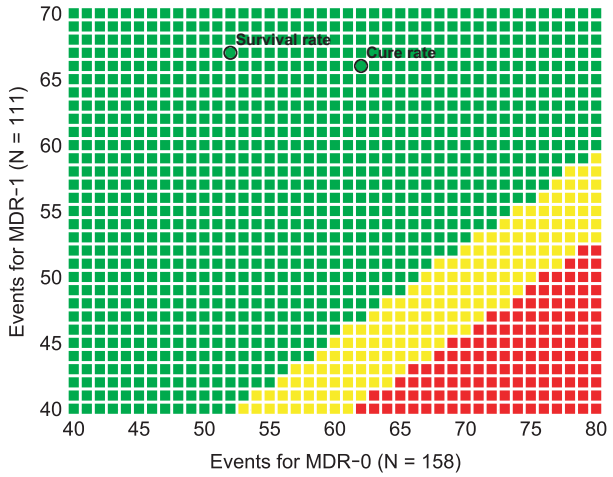

Fig. 4. Data-driven scoring based on confidence interval scoring. Scoring table for number of survivors and cured patients, based on the estimates from fig. 2. The colour red indicates that MDR-1 is inferior, yellow non-inferiority and green superiority. MDR-1 shows a superior profile with regard to both the number of cured (59% *versus* 39%) and survived (60% *versus* 33%) patients.

For myocardial ischaemia, the MDR-1 group has an inferior profile with three cases of myocardial ischaemia *versus* 0 cases in MDR-0. The MDR-1 group is also inferior on the infection criterion. For the remaining criteria, the MDR-1 group is non-inferior. Figure 4 shows that patients in the MDR-1 group have a clear superior profile with regard to cure and survival rate.

Additional results using common statistical methods were also performed. TTR and TTD were assessed using univariable and multivariable Cox proportional hazards models. Adjustment was made for known predictive factors (age, sex, tumour grade and tumour stage) and cohort in the multivariable Cox proportional hazards models. The results showed that the hazard ratio (CI 95%) for MDR-1 with respect to TTR is 0.66 [0.45–0.95]; $p = 0.03$, where MDR-0 is defined as HR = 1. The MDR classifier showed a similar tendency but was not significant for TTD where the hazard ratio was 0.70 [0.47–1.02]; $p = 0.06$.

The results for 'survival rate' and 'cure rate' were statistically significant at a 95% confidence level using a Fischer's exact test (two-tailed test), while the remaining criteria were not significant. A Barnard's exact test (two-tailed test) showed that, additionally, the result for 'myocardial ischaemia' was statistically significant at a 95% confidence level, while the remaining again were not significant. At the usual level of confidence (95%), only cure and survival rate were significant, while the rest of the criteria were not significant using Clopper–Pearson's exact method.

Evidence and uncertainty.

The scores that are calculated in step 4 – Scoring – and shown in table 4 must be reviewed critically [3]. The scores are based on a total of 269 patients, distributed on the two groups. The basis for the scores is relatively weak, because of the low number of patients, and this is accounted for in the data-driven scoring method in a simple way: The scores of borderline criteria are changed to intervals. By borderline criteria, we mean that a change in the number of events by only one would change the score. Arthralgia/myalgia, infections and fatigue are all borderline, and the scores are consequently changed to intervals, as seen in table 4. Time-to-death curves show that 64% of patients dying in the MDR-1 group have a longer time-to-death, while the difference distribution for time-to-relapse shows that 55% of the patients surviving in the MDR-1 group have longer time-to-relapse relative to the surviving patients in the MDR-0 group. According to the definition of clinical relevance, the MDR-1 profile is not superior to the MDR-0 profile with regard to time-to-relapse, but there is a tendency in favour of MDR-1. However, this difference is regarded as clinically not significant.

For the criteria infections, arthralgia/myalgia and fatigue, only one more case in one of the groups will shift the score. Based on these observations, the scores for these criteria are represented by intervals.

Table 4.

Criteria, weights, scores and weighted scores.

Criterion	Weight	Scores MDR-0	Scores MDR-1	Weighted scores MDR-0	Weighted scores MDR-1
Cure rate	3	−1	1	−3	3
Survival rate	3	−1	1	−3	3
TTD	3	−1	1	−3	3
TTR	3	0	0	0	0
Infection	2	1→0	−1→0	2→0	−2→0
Myocardial ischaemia	2	1	−1	2	−2
Bleeding	2	0	0	0	0
Mucositis/stomatitis	2	−1	1	−2	2
Hand-foot skin syndrome	2	0	0	0	0
Diarrhoea	2	0	0	0	0
Arthralgia/myalgia	1	1→0	−1→0	1→0	−1→0
Fatigue	1	−1→0	0→1	−1→0	0→1
Nausea/vomiting	1	0	0	0	0

TTD, time-to-death; TTR, time-to-relapse.

Weighted scores.

The weights and scores are multiplied to produce weighted scores that enable the direct comparison of different criteria between the two options. The weighted scores capture both the importance and performance. Table 4 shows criteria, their weights, scores and weighted scores.

Visualisation.

The results of the assessment are visualised and communicated in a tornado-like diagram, as seen in fig. 5, where the *weight* of a criterion is depicted as the width of box and the *score* is represented by the colour. The wider the box, the more important the criterion is. If the MDR-1 is superior, the colour green is used, yellow if there is no difference and red if the MDR-1 is inferior. Figure 5 shows that patients in MDR-1 have an equivalent or better outcome for the most important criteria, cure rate, survival rate, TTD and TTR, while only inferior to non-inferior for one medium important and one low important criterion, infections and arthralgia/myalgia, respectively. In the tornado diagram, both of these are marked as inferior (red). Patients in the MDR-1 group have clearly inferior profile for only one criterion of medium importance, myocardial ischaemia.

Overall conclusion of the assessment.

A clinically significant and relevant difference for the high-importance criteria, cure rate, survival rate and TTD was found in favour of the MDR-1 group, while for TTR, the results were non-inferior, but a tendency in favour of the MDR-1 group is seen. A higher risk of severe cases of the medium importance criterion myocardial ischaemia and a slightly higher risk of the medium importance criterion infection were seen in MDR-1, which needs to be studied. Because of the limited number of patients, the results for the criteria TTR, infections, arthralgia/myalgia and fatigue are not conclusive and are only seen as trends/tendencies.

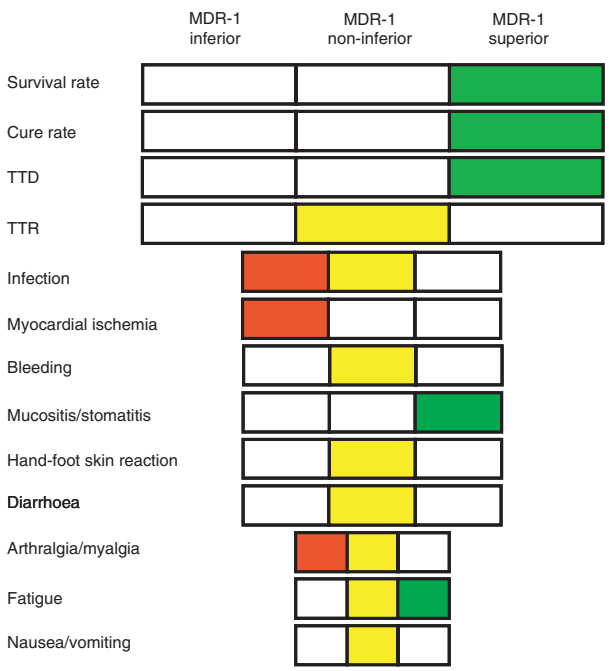

Fig. 5. Tornado-like diagram used in visualisation. The colour red indicates that MDR-1 is inferior, yellow non-inferiority and green superiority. The results of the assessment are shown, where patients in the MDR-1 group have superior profile with regard to three of four most important criteria, survival rate, cure rate and time-to-death, while non-inferior for TTR.

The clinical implications of this study are that the genetic profile of the MDR-1 group has a clinically substantially superior profile with regard to three of four most important criteria, cure rate, survival rate and TTD and probably also with regard to TTR, which is the last of the most important criteria. Furthermore, patients in the MDR-1 group have a superior profile for the medium important criterion mucositis/stomatitis and probably superior to non-inferior on the low importance criterion fatigue. Patients in MDR-0 have a superior profile for myocardial ischaemia and also probably superior profile for infections and arthralgia/myalgia.

The combination of genetic polymorphisms for patients in the MDR-1 group in comparison with the MDR-0 group causes clinically relevant differences for several clinical aspects, and the overall benefit–risk profile is positive. Based on this study, genetic profiling is therefore advisable in patients with colorectal cancer, to enable individualised treatment and follow-up.

Discussion

Statistical significant results are essential, but not on expense of the clinical relevance of data [12–14]. Statistical significance cannot be used to conclude that an intervention is clinically relevant. We propose that clinical relevance is defined and justified based on the proportion of patients experiencing a specified treatment effect, and we have demonstrated how clinical relevance can be quantified and discussed in a qualitative manner. However, as mentioned previously, the proposed methodology is not to be seen as a substitute of existing statistical analysis but a way to place the results of a statistical analysis in a broader perspective.

Common statistical tests were performed at 95% confidence level. Hazard ratios were used for TTD and TTR, while Fisher's exact test and Barnard's exact test were used for the remaining. The results show that TTR is statistically significant using hazard ratio, while cure and survival rate are statistically significant using Fisher's, Barnard's and Clopper–Pearson's exact tests. Additionally, myocardial ischaemia was found to be statistically significant under Barnard's exact test. However, looking at fig. 3, it is seen that there is a clear trend in favour of the MDR-1 group. The suggested 2/3 threshold has proved to be reasonable, because the balance between type I and II errors is acceptable.

While supported by statistical analysis, the proposed method itself is of a qualitative nature to properly allow for uncertainties and differences in opinion. This qualitative feature also serves to focus the assessment on clinical and toxicological issues, and it allows comparison of benefits and risks that have largely different probability and/or significance in a justifiable detail.

The difference distribution captures even the smallest change or difference. One can discuss whether or not patients surviving a day or a week more should be taken into account. If wanted, a cut-off point could be defined enabling, for example, only patients that survive more than 1 month to be accounted for. However, it might be difficult to pre-define such a value but this is precisely one of the main advantages of the proposed methodology: decision makers are forced to conduct such valuable and important discussions prior to and after the data analysis. This aspect of method can be tailored to the specific scenario, situation and decision context.

The method allows comparison of different criteria that have largely different probability and/or significance in a justifiable detail, for example, rare events of lower importance with continuous data of higher importance such as time-to-death. The results show that patients in the MDR-1 group clearly have an inferior profile with regard to severe cases of the medium importance criterion myocardial ischaemia, while patients in the MDR-0 group clearly have an inferior profile with regard to severe cases of mucositis/stomatitis.

The weights applied to the individual criteria can obviously be questioned and discussed. Ideally, weights are to be decided in consensus by clinicians, regulators and patient organisations. The idea of the overall procedures is that different stakeholders can locate and evaluate their possible points of divergence. This may lead to an adjustment of the weights through more detailed argumentation.

Different stakeholders are likely to have different views on weighting and scoring of the various criteria, and they may also disagree about the choice of criteria. It is therefore important that all relevant information is maintained and clearly presented in the final decision process.

In our experience, the structured process enables valuable discussions in a structured way. It is these valuable and structured discussions that create the foundation for creditable assessments. The results can then easily be communicated in standardised diagrams, where essential information about importance and performance is visualised.

In conclusion, we have demonstrated a comprehensive approach to data-driven assessments and how it can be used in a clinical setting. The method can handle a variety of different types of clinical data and can be used in single trials as well as in multiple trials [15]. Preliminary tests of the methodology have given satisfactory results, particularly concerning the ability of the method to retain the most relevant information for the final assessment. The proposed methodology helps different stakeholders to compare their views on a point-to-point basis in a transparent and structured way with focus on the clinical relevance of data. Based on this study, genetic profiling is advisable in patients with colorectal cancer, to enable individualised treatment and follow-up. Further studies are, obviously, needed to consolidate these findings.

Conflict of Interest/Disclosure

The views and opinions expressed in this study are those of the individual authors and should not be attributed to Novo Nordisk A/S, Danish Medicines Agency, Technical University of Denmark, Rigshospitalet (Copenhagen University Hospital) or Bispebjerg Hospital. S.B. Sarac and C.H. Rasmussen are at the time of submission employed as full-time industrial PhD students at the Technical University of Denmark and Novo Nordisk A/S. S. Thirstrup is a full-time employee at the Danish Medicines Agency, a governmental organisation. No funding was received for this study.

References

1 The Council for International Organizations of Medical Sciences (CIOMS). Working CIOMS group IV – benefit-risk balance for marketed drugs: evaluating safety signals [online]. Available from URL: http://www.cioms.ch/publications/g4-benefit-risk.pdf. 1998.
2 US Food and Drug Administration (FDA). The future of drug safety – promoting and protecting the health of the public [online]. Available from URL: http://www.fda.gov/downloads/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/UCM171627.pdf. 2007.
3 Committee for Medicinal Products for Human Use (CHMP). Report of the CHMP working group on benefit-risk assessment models and methods. EMEA/CHMP/15404/2007 [online]. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/01/WC500069634.pdf. CHMP; 2008.
4 Measuring Benefit and Balancing Risk: Strategies for the benefit-risk assessment of new medicines in a risk-averse environment Washington, DC, US: 19–20 June 2008. Report of the Workshop organised by the CMR International Institute for Regulatory Science. Available from URL: http://www.cmr.org. 2008.
5 Strategies for Communicating Benefit-Risk to Decision Makers: Explaining Methods, Findings and Conclusions Through a Common Approach Washington, DC, US: 17–19 June 2009. Report of the Workshop organised by the CMR International Institute for Regulatory Science. Available from URL: http://www.cmr.org/workshops. 2009.
6 Benefit-risk methodology project – work package 2 report: applicability of current tools and processes for regulatory benefit-risk assessment. EMA/549682/2010[online]. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/10/WC500097750.pdf. European Medicines Agency (EMA); 2010.
7 Benefit risk methodology project – work package 1: description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network. EMA/213482/2010 [online]. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/04/WC500089603.pdf. 2010.
8 Mussen F, Salek S, Walker S. 978-0-470-06085-8 (H/B) Benefit-Risk Appraisal of Medicines – A Systematic Approach to Decision-Making. Wiley-Blackwell, West Sussex, UK, 2009.
9 Walker S, McAuslane N, Liberti L, Salek S. Measuring benefit and balancing risk: strategies for the benefit-risk assessment of new medicines in a risk-averse environment. Clin Pharmacol Ther 2009;**85**:241–6.
10 Walker S, Liberti L, McAuslane N. Refining the benefit-risk framework for the assessment of medicines: valuing and weighting benefit and risk parameters. Clin Pharmacol Ther 2011;**89**:179–82.
11 Nutt DJ, King LA, Phillips L. Drug harms in the UK: a multicriteria decision analysis. Lancet 2010;**376**:1558–65.
12 Greenstein G. Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. J Am Dent Assoc 2003;**134**:583–91.

13 Boudes PF. How to improve complex drug development? A critical review of FDA advisory meetings. Drug Inf J 2007;**41**:673–83.
14 McCloskey DN, Ziliak ST. The unreasonable ineffectiveness of fisherian "tests" in biology and especially medicine. Biological Theory 2009;**4**:44–53.
15 Sarac SB, Rasmussen CH, Rasmussen MA, Hallgreen CE, Søeborg T, Colding-Jørgensen M *et al.* A comprehensive approach to benefit-risk assessment in drug development. Basic Clin Pharmacol Toxicol 2012; doi:10.1111/j.1742-7843.2012.00871.x [Epub ahead of print].
16 Afzal S, Gusella M, Jensen SA, Vainer B, Vogel U, Andersen JT *et al.* The association of polymorphisms in 5-fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. Pharmacogenomics 2011;**12**:1257–67.
17 Danenberg PV, Danenberg KD. Effect of 5, 10-methylenetetrahydrofolate on the dissociation of 5-fluoro-2'-deoxyuridylate from thymidylate synthetase: evidence for an ordered mechanism. Biochemistry 1978;**17**:4018–24.
18 Armstrong RD, Takimoto CH, Cadman EC. Fluoropyrimidine-mediated changes in small nuclear RNA. J Biol Chem 1986; **261**:21–4.
19 Doong SL, Dolnick BJ. 5-Fluorouracil substitution alters pre-mRNA splicing in vitro. J Biol Chem 1988;**263**:4467–73.
20 Houghton JA, Maroda SJ Jr, Phillips JO, Houghton PJ. Biochemical determinants of responsiveness to 5-fluorouracil and its derivatives in xenografts of human colorectal adenocarcinomas in mice. Cancer Res 1981;**41**:144–9.
21 Beck A, Etienne MC, Cheradame S, Fischel JL, Formento P, Renee N, *et al.* A role for dihydropyrimidine dehydrogenase and thymidylate synthase in tumour sensitivity to fluorouracil. Eur J Cancer 1994;**30A**:1517–22.
22 van Triest B, Pinedo HM, van Hensbergen Y, Smid K, Telleman F, Schoenmakers PS *et al.* Thymidylate synthase level as the main predictive parameter for sensitivity to 5-fluorouracil, but not for folate-based thymidylate synthase inhibitors, in 13 nonselected colon cancer cell lines. Clin Cancer Res 1999;**5**:643–54.
23 Takebe N, Zhao SC, Ural AU, Johnson MR, Banerjee D, Diasio RB *et al.* Retroviral transduction of human dihydropyrimidine dehydrogenase cDNA confers resistance to 5-fluorouracil in murine hematopoietic progenitor cells and human CD34 + -enriched peripheral blood progenitor cells. Cancer Gene Ther 2001;**8**:966–73.
24 Etienne MC, Ilc K, Formento JL, Laurent-Puig P, Formento P, Cheradame S, *et al.* Thymidylate synthase and methylenetetrahydrofolate reductase gene polymorphisms: relationships with 5-fluorouracil sensitivity. Br J Cancer 2004;**90**:526–34.
25 Sohn KJ, Croxford R, Yates Z, Lucock M, Kim YI. Effect of the methylenetetrahydrofolate reductase C677T polymorphism on chemosensitivity of colon and breast cancer cells to 5-fluorouracil and methotrexate. J Natl Cancer Inst 2004;**96**:134–44.
26 Sakamoto E, Nagase H, Kobunai T, Oie S, Oka T, Fukushima M *et al.* Orotate phosphoribosyltransferase expression level in tumors is a potential determinant of the efficacy of 5-fluorouracil. Biochem Biophys Res Commun 2007;**363**:216–22.
27 Edler D, Glimelius B, Hallstrom M, Jakobsen A, Johnston PG, Magnusson I *et al.* Thymidylate synthase expression in colorectal cancer: a prognostic and predictive marker of benefit from adjuvant fluorouracil-based chemotherapy. J Clin Oncol 2002;**20**:1721–8.
28 Kornmann M, Schwabe W, Sander S, Kron M, Strater J, Polat S *et al.* Thymidylate synthase and dihydropyrimidine dehydrogenase mRNA expression levels: predictors for survival in colorectal cancer patients receiving adjuvant 5-fluorouracil. Clin Cancer Res 2003;**9**:4116–24.
29 Tsuji T, Hidaka S, Sawai T, Nakagoe T, Yano H, Haseba M *et al.* Polymorphism in the thymidylate synthase promoter enhancer region is not an efficacious marker for tumor sensitivity to 5-fluorouracil-based oral adjuvant chemotherapy in colorectal cancer. Clin Cancer Res 2003;**9**(10 Pt 1):3700–4.
30 Popat S, Matakidou A, Houlston RS. Thymidylate synthase expression and prognosis in colorectal cancer: a systematic review and meta-analysis. J Clin Oncol 2004;**22**:529–36.
31 Hitre E, Budai B, Adleff V, Czegledi F, Horvath Z, Gyergyay F *et al.* Influence of thymidylate synthase gene polymorphisms on the survival of colorectal cancer patients receiving adjuvant 5-fluorouracil. Pharmacogenet Genomics 2005;**15**:723–30.
32 Dotor E, Cuatrecases M, Martinez-Iniesta M, Navarro M, Vilardell F, Guino E *et al.* Tumor thymidylate synthase 1494del6 genotype as a prognostic factor in colorectal cancer patients receiving fluorouracil-based adjuvant treatment. J Clin Oncol 2006;**24**:1603–11.
33 Ochiai T, Nishimura K, Noguchi H, Kitajima M, Tsuruoka Y, Takahashi Y *et al.* Prognostic impact of orotate phosphoribosyl transferase activity in resectable colorectal cancers treated by 5-fluorouracil-based adjuvant chemotherapy. J Surg Oncol 2006; **94**:45–50.
34 Terrazzino S, Agostini M, Pucciarelli S, Pasetto LM, Friso ML, Ambrosi A *et al.* A haplotype of the methylenetetrahydrofolate reductase gene predicts poor tumor response in rectal cancer patients receiving preoperative chemoradiation. Pharmacogenet Genomics 2006;**16**:817–24.
35 Lurje G, Zhang W, Yang D, Groshen S, Hendifar AE, Husain H *et al.* Thymidylate synthase haplotype is associated with tumor recurrence in stage II and stage III colon cancer. Pharmacogenet Genomics 2008;**18**:161–8.
36 Soong R, Shah N, Salto-Tellez M, Tai BC, Soo RA, Han HC *et al.* Prognostic significance of thymidylate synthase, dihydropyrimidine dehydrogenase and thymidine phosphorylase protein expression in colorectal cancer patients treated with or without 5-fluorouracil-based chemotherapy. Ann Oncol 2008;**19**:915–9.
37 Afzal S, Jensen SA, Vainer B, Vogel U, Matsen JP, Sorensen JB *et al.* MTHFR polymorphisms and 5-FU-based adjuvant chemotherapy in colorectal cancer. Ann Oncol 2009;**20**:1660–6.
38 Gusella M, Frigo AC, Bolzonella C, Marinelli R, Barile C, Bononi A *et al.* Predictors of survival and toxicity in patients on adjuvant therapy with 5-fluorouracil for colorectal cancer. Br J Cancer 2009;**100**:1549–57.
39 Zintzaras E, Ziogas DC, Kitsios GD, Papathanasiou AA, Lau J, Raman G. MTHFR gene polymorphisms and response to chemotherapy in colorectal cancer: a meta-analysis. Pharmacogenomics 2009;**10**:1285–94.
40 Park CM, Lee WY, Chun HK, Cho YB, Yun HR, Heo JS *et al.* Relationship of polymorphism of the tandem repeat sequence in the thymidylate synthase gene and the survival of stage III colorectal cancer patients receiving adjuvant 5-flurouracil-based chemotherapy. J Surg Oncol 2010;**101**:22–7.
41 Afzal S, Gusella M, Vainer B, Vogel UB, Andersen JT, Broedbaek K *et al.* Combinations of polymorphisms in genes involved in the 5-Fluorouracil metabolism pathway are associated with gastrointestinal toxicity in chemotherapy-treated colorectal cancer patients. Clin Cancer Res 2011;**17**:3822–9.
42 Keeney RL. Value Focused Thinking – A Path to Creative Decisionmaking. Harvard University Press, Massachusetts, USA, 1992.
43 Grinstead CM, Snell JL. Sums of Random Variables. Introduction to Probability, 2nd edn. American Mathematical Society, 1997;285.
44 Clopper C, Pearson E. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;**26**: 404–13.
45 Goetghebeur MM, Wagner M, Khoury H, Levitt RJ, Erickson LJ, Rindress D. Evidence and value: impact on decisionmaking–the EVIDEM framework and potential applications. BMC Health Serv Res 2008;**8**:270.