Table 2. Variations of amino acids in NS5A region from 33 patients before IFN therapy

*N terminus(1973-2118)

Position	HCV-J	Amino acid	Outcome	Ratio
2024	I	I	SVR	6/11
			ETR	8/11
			NR	4/11
		V	SVR	5/11
			ETR	3/11
			NR	7/11

* Central(2119-2280)

Position	HCV-J	Amino acid	Outcome	Ratio
2125	V	V	SVR	4/11
			ETR	7/11
			NR	4/11
		L	SVR	7/11
			ETR	4/11
			NR	7/11
2155	L	A	SVR	0/11
			ETR	1/11
			NR	0/11
		V	SVR	2/11
			ETR	3/11
			NR	0/11
		P	SVR	8/11
			ETR	3/11
			NR	9/11
		L	SVR	1/11
			ETR	4/11
			NR	2/11
2218	H	H	SVR	3/11
			ETR	8/11
			NR	6/11
		R	SVR	6/11
			ETR	3/11
			NR	3/11
		C	SVR	1/11
			ETR	0/11
			NR	1/11
		N	SVR	0/11
			ETR	0/11
			NR	1/11
		F	SVR	1/11
			ETR	0/11
			NR	0/11
2271	P	P	SVR	3/11
			ETR	6/11
			NR	3/11
		A	SVR	7/11
			ETR	4/11
			NR	6/11
		E	SVR	1/11
			ETR	1/11
			NR	2/11
2278	P	S	SVR	4/11
			ETR	4/11
			NR	2/11
		T	SVR	6/11
			ETR	4/11
			NR	8/11
		P	SVR	1/11
			ETR	3/11
			NR	1/11

* C terminus(2281-2419)

Position	HCV-J	Amino acid	Outcome	Ratio
2321	A	A	SVR	8/11
			ETR	7/11
			NR	10/11
		G	SVR	0/11
			ETR	2/11
			NR	0/11
		T	SVR	2/11
			ETR	1/11
			NR	1/11
		V	SVR	1/11
			ETR	1/11
			NR	0/11
2334	V	V	SVR	6/11
			ETR	10/11
			NR	5/11
		I	SVR	5/11
			ETR	1/11
			NR	6/11
2356	G	E	SVR	1/11
			ETR	7/11
			NR	2/11
		G	SVR	8/11
			ETR	4/11
			NR	8/11
		Q	SVR	0/11
			ETR	0/11
			NR	1/11
		R	SVR	1/11
			ETR	0/11
			NR	0/11
		T	SVR	1/11
			ETR	0/11
			NR	0/11
2378	K	A	SVR	5/11
			ETR	5/11
			NR	0/11
		T	SVR	3/11
			ETR	1/11
			NR	7/11
		K	SVR	2/11
			ETR	5/11
			NR	4/11
		P	SVR	1/11
			ETR	0/11
			NR	0/11

Numbering of amino acid was based on that of HCV-J strain.

Table 3. Relationship between amino acid at position 2378 and the treatment outcome of IFN

Treatment outcome	Amino acid in 2378				Total
	A*	T[#]	K	P	
SVR	5	3	2	1	11
ETR	5	1	5	0	11
NR	0	7	4	0	11
Total	**10**	**11**	**11**	**1**	**33**
%	**30.3**	**33.3**	**33.3**	**3.0**	**100**

*P=0.022 (Alanine versus non-Alanine).

[#] P=0.028 (Threonine versus non-Threonine).

Abbrevations: SVR, sustained virological responder; ETR, end-of-treatment responder; NR, non-responder; A, alanine; T, threonine; K, lysine; P, proline.

Table 4. Relationship between amino acid at position 2378 and the number of mutations in ISDR

Mutations in ISDR	Treatment outcome	Amino acid in 2378				**Total**
		A	T	K	P	
ISDR>=4	SVR	4	1	2	1	8
	ETR	0	0	1	0	1
	NR	0	1	2	0	3
total		**4**	**2**	**5**	**1**	**12**
%		**33.3**	**16.7**	**41.7**	**8.3**	**100**
ISDR<4	SVR	1	2	0	0	3
	ETR	5	1	4	0	10
	NR	0	6	2	0	8
Total		**6**	**9**	**6**	**0**	**21**
%		**28.6**	**42.9**	**28.6**	**0**	**100**

Abbrevetions: ISDR, Interferon sensitivity determining region;
A, alanine; T, threonine; K, lysine; P, proline.

Journal of Medical Virology

Sequences in the Interferon Sensitivity Determining Region and Core Region of Hepatitis C Virus Impact Pretreatment Prediction of Response to Peg-interferon Plus Ribavirin: Data Mining Analysis

Journal:	*Journal of Medical Virology*
Manuscript ID:	JMV-10-2053.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kurosaki, Masayuki; Musashino Red Cross Hospital Sakamoto, Naoya; Tokyo Medical and Dental University, Dept. of Gastroenterology and Hepatology Iwasaki, Manabu Sakamoto, Minoru Suzuki, Yoshiyuki Hiramatsu, Naoki; Osaka University School of Medicine, Department of Gastroenterology and Hepatology Sugauchi, Fuminaka; Nagoya city university, Graduate school of medical science, Depatrment of Gastroenterology and Metabolism Tamori, Akihiro; Osaka City University Gradute Shool of Medicine, Hepatoogy Nakagawa, Mina Izumi, Namiki; Musashino Red Cross Hospital, Division of Gastroenterology and Hepatology
Keywords:	ISDR, data mining, decision tree

Sequences in the Interferon Sensitivity Determining Region and Core Region of Hepatitis C Virus Impact Pretreatment Prediction of Response to Peg-interferon Plus Ribavirin: Data Mining Analysis

Short title: Prediction of response to PEG-interferon RBV therapy

Masayuki Kurosaki[1], Naoya Sakamoto[2], Manabu Iwasaki[3], Minoru Sakamoto[4], Yoshiyuki Suzuki[5], Naoki Hiramatsu[6], Fuminaka Sugauchi[7], Akihiro Tamori[8], Mina Nakagawa[2], and Namiki Izumi[1]

[1]Division of Gastroenterology and Hepatology, Musashino Red Cross Hospital, Tokyo, Japan

[2]Department of Gastroenterology and Hepatology, Tokyo Medical and Dental University, Tokyo, Japan

[3]Department of Computer and Information Science, Seikei University, Tokyo, Japan

[4]First Department of Internal Medicine, University of Yamanashi, Yamanashi, Japan

[5]Department of Hepatology, Toranomon Hospital, Tokyo, Japan

[6]Department of Gastroenterology and Hepatology, Osaka University Graduate School of

Medicine, Osaka, Japan

[7]Department of Gastroenterology and Metabolism, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan

[8]Department of Hepatology, Osaka City University Medical School, Osaka, Japan

Grant Support: This study was supported by a Grant-In-Aid from the Ministry of Health, Labor and Welfare, Japan.

Conflicts of interest: No conflicts of interest exist for all authors.

Word count: 3595 words

Address correspondence to:

Namiki Izumi, M.D.

Division of Gastroenterology and Hepatology,

Musashino Red Cross Hospital

1-26-1 Kyonan-cho, Musashino-shi, Tokyo 180-8610, Japan

Tel: +81-422-32-3111

Fax: +81-422-32-9551

E-mail: nizumi@musashino.jrc.or.jp

List of authors

Masayuki Kurosaki	kurosaki@musashino.jrc.or.jp
Naoya Sakamoto	nsakamoto.gast@tmd.ac.jp
Manabu Iwasaki	iwasaki@st.seikei.ac.jp
Minoru Sakamoto	msakamoto@yamanashi.ac.jp
Yoshiyuki Suzuki	suzunari@seagreen.ocn.ne.jp
Naoki Hiramatsu	hiramatsu@gh.med.osaka-u.ac.jp
Fuminaka Sugauchi	fsugauch@med.nagoya-cu.ac.jp
Akihiro Tamori	atamori@med.osaka-cu.ac.jp
Mina Nakagawa	mnakagawa.gast@tmd.ac.jp
Namiki Izumi	nizumi@musashino.jrc.or.jp

Abstract

The aim of the present study is to clarify the significance of viral factors for pretreatment prediction of sustained virological response to peg-interferon (PEG-IFN) plus ribavirin (RBV) therapy for chronic hepatitis C using data mining analysis. Substitutions in the IFN sensitivity-determining region (ISDR) and at position 70 of the HCV core region (Core70) were determined in 505 patients with genotype 1b chronic hepatitis C treated with PEG-IFN plus RBV. Data mining analysis was used to build a predictive model of sustained virological response in patients selected randomly (n = 304). The reproducibility of the model was validated in the remaining 201 patients. Substitutions in ISDR (odds ratio = 9.92, $p < 0.0001$) and Core70 (odds ratio = 1.92, $p = 0.01$) predicted sustained virological response independent of other covariates. The decision-tree model revealed that the rate of sustained virological response was highest (83%) in patients with 2 or more substitutions in ISDR. The overall rate of sustained virological response was 44% in patients with a low number of substitutions in ISDR (0–1) but was 83% in selected subgroups of younger patients (<60 years), wild-type sequence at Core70, and higher level of low-density lipoprotein cholesterol (LDL-C) (≥120 mg/dL). Reproducibility of the model was validated ($r^2 = 0.94$, $p < 0.001$). In conclusion, substitutions in ISDR and Core70 of HCV are significant predictors of response to PEG-IFN plus RBV therapy. A decision-tree model that includes these viral factors as predictors could identify patients efficiently with a high probability of sustained virological response.

Key words: data mining, decision-tree model, ISDR, core region, peg-interferon

Introduction

The combination of pegylated-interferon (PEG-IFN) plus ribavirin (RBV) is currently the most effective therapy for chronic hepatitis C, but the rate of sustained virological response after 48 weeks of therapy is about 50% in patients with HCV genotype 1b and a high HCV RNA titer [Fried et al., 2002; Manns et al., 2001]. The most reliable means to predict sustained virological response is to monitor the viral response during the early weeks of treatment. The early virological response, defined as undetectable HCV RNA at week 12, is associated with a high rate of sustained virological response [Davis et al., 2003; Lee and Ferenci, 2008]. The rapid virological response, defined as undetectable HCV RNA at week 4 of therapy, is even more predictive of sustained virological response than the early virological response [Izumi et al., 2010; Jensen et al., 2006; Yu et al., 2008]. However, there is no established means that predicts the virological response before commencing treatment. Recent reports have revealed that single nucleotide polymorphisms located near the *IL28B* gene show a strong association with the response to PEG-IFN plus RBV therapy [Ge et al., 2009; Kurosaki et al., 2010c; Suppiah et al., 2009; Tanaka et al., 2009]. These findings indicate that the host factor is an important determinant of the treatment response. On the other hand, the present study's authors have reported that a stretch of 40 amino acids in the NS5A region of HCV, designated as the interferon sensitivity-determining region (ISDR), has a close association with the virological response to interferon mono-therapy [Enomoto et al., 1995; Enomoto et al., 1996; Kurosaki et al., 1997]. More recently, amino acid substitutions at positions 70 and 91 of the core region have been reported to be associated with response to PEG-IFN plus RBV combination therapy [Akuta et al., 2007a; Akuta et al., 2005]. The impact of these HCV substitutions on treatment response is yet to be validated.

Decision-tree analysis is a core component of data mining analysis that can be used to build predictive models [Breiman et al., 1980]. This method has been used to define prognostic factors in various diseases such as prostate cancer [Garzotto et al., 2005], diabetes [Miyaki et al., 2002], melanoma [Averbook et al., 2002; Leiter et al., 2004], colorectal carcinoma [Valera et al., 2007; Zlobec et al., 2005], and liver failure [Baquerizo et al., 2003]. The major advantage of decision-tree analysis over logistic regression analysis is that the results of analysis are easy to understand. The simple allocation of patients into subgroups by following the flowchart form could define the predicted possibility of outcome [LeBlanc and Crowley, 1995].

The present study's authors used decision-tree analysis for the prediction of early virological response (undetectable HCV RNA within 12 weeks of therapy) to

PEG-IFN and RBV combination therapy in chronic hepatitis C [Kurosaki et al., 2010a], and more recently for the pretreatment prediction of sustained virological response [Kurosaki et al., 2010b]. In the latter model, simple and noninvasive standard tests were used as parameters; specialized tests such as viral mutations and host genetics, or invasive tests such as liver histology, were not included because the aim of that model was for use in general medical practice, especially in some countries or areas where resources are limited. Thus, the impact of viral mutations or liver histology was not considered in that model.

The present study analyzed whether including viral substitutions in ISDR and the core region of HCV in the decision-tree model could improve its predictive accuracy over the previous model to identify chronic hepatitis C patients who are likely to respond to PEG-IFN plus RBV therapy.

Materials and methods

Patients

This multicenter retrospective cohort study included 505 chronic hepatitis C patients who were treated with PEG-IFN alpha-2b and RBV at Musashino Red Cross Hospital, Toranomon Hospital, Tokyo Medical and Dental University, Osaka University, Nagoya City University Graduate School of Medical Sciences, Yamanashi University, Osaka City University, and their related hospitals. The inclusion criteria were: (1) genotype 1b, (2) HCV RNA titer higher than 100 KIU/mL by quantitative PCR (Cobas Amplicor HCV Monitor v 2.0, Roche Diagnostic Systems, CA), (3) no co-infection with hepatitis B virus or human immunodeficiency virus, (4) no other causes of liver disease, (5) patients having undergone liver biopsy prior to IFN treatment, (6) number of substitutions in ISDR having been determined, (7) substitutions in the amino acid positions 70 and 91 of the core region having been determined, and (8) completion of at least 12 weeks of therapy. Patients were treated with PEG-IFN alpha-2b (1.5 μg/kg) weekly plus RBV. The daily dose of RBV was adjusted by weight: 600 mg for <60 kg, 800 mg for 60–80 kg, and 1000 mg for >80 kg. For the analysis, patients were assigned randomly to either the model building (304 patients) or validation (201 patients) groups. There were no significant differences in the clinical backgrounds between these 2 groups (Table 1). Informed consent was obtained from each patient. The study protocol conformed to the ethical guidelines of the Declaration of Helsinki and was approved by the institutional review committees of all concerned hospitals.

Laboratory tests

Hematological tests, blood chemistry, and HCV RNA titer were analyzed before therapy and at least once every month during therapy. Sequences of ISDR and the core region of HCV were determined by direct sequencing after amplification by reverse transcription and polymerase chain reaction as reported previously. At position 70 of the core region (Core70), arginine was defined as the wild type, and glutamine or histidine was defined as the mutant type. At position 91 of the core region, leucine was defined as the wild type and methionine was defined as the mutant type, as described previously [Akuta et al., 2005]. Fibrosis and activity were scored according to the METAVIR scoring system [Bedossa and Poynard, 1996]. Fibrosis was staged on a scale of 0–4: F0 (no fibrosis), F1 (mild fibrosis), F2 (moderate fibrosis), F3 (severe fibrosis) and F4 (cirrhosis). Activity of necroinflammation was graded on a scale of 0–3: A0 (no activity), A1 (mild activity), A2 (moderate activity), and A3 (severe activity). Sustained virological response was defined as undetectable HCV RNA by qualitative

PCR with a lower detection limit of 50 IU/mL (Amplicor, Roche Diagnostic systems, CA) at week 24 after the completion of therapy.

Statistical analysis

A database of pretreatment variables included hematological tests (hemoglobin level, white blood cell count, and platelet count), blood chemistry tests (serum levels of creatinine, albumin, aspartate aminotransferase, alanine aminotransferase (ALT), gamma-glutamyltransferase (GGT), total cholesterol, triglyceride, and low-density lipoprotein cholesterol (LDL-C)), viral factors (HCV RNA titer, number of substitutions in ISDR, substitutions in the amino acid positions 70 and 91 of the core region), histological findings (stage of fibrosis and grade of activity) and patient characteristics (age, sex, and body mass index). Based on this database, decision-tree analysis was used to define a predictive model for sustained virological response.

Student's t test was used for the univariable comparison of quantitative variables and Fisher's exact test was used for the comparison of qualitative variables. For the multivariable analysis for factors associated with sustained virological response, logistic regression models with backward selection were used to identify independent predictors of sustained virological response. Variables that showed significant association with sustained virological response by univariable analysis were included in the multivariable analysis. IBM-SPSS software v.15.0 (SPSS Inc, Chicago, IL) was used for these analyses. For the decision-tree analysis [Segal and Bloch, 1989], the data mining software IBM SPSS Modeler 13 (IBM SPSS Inc, Chicago, IL) was used, as reported previously [Kurosaki et al., 2010a; Kurosaki et al., 2010b]. In brief, the software searched for the optimal split variables to build a decision-tree structure. The entire study population was first evaluated to determine the variables and cut-off points for the most significant division into 2 subgroups having different probabilities of sustained virological response. Thereafter, analysis was repeated on all subgroups in the same way until either no additional significant variable was detected or the sample size was below 20.

Results

Generation of the Decision-Tree model

The decision-tree analysis selected 5 predictive variables to produce 6 subgroups of patients (Figure 1). The number of substitutions in ISDR was selected as the best predictor of sustained virological response. The possibility of achieving sustained virological response was 83% for patients with 2 or more substitutions in ISDR compared with 44% for patients with a single or no substitution. Among patients with a single or no substitution in ISDR, age, with an optimal cut-off of 60 years, was selected as the variable of second split. Patients younger than 60 had the higher probability of sustained virological response (55%) compared with those older than 60 years (31%). Among younger patients, amino acid substitution at Core70 was selected as the third variable of split—wild-type sequence being the predictor of favorable response compared with the mutant type (65% versus 36%). Among patients with wild-type Core70, the level of serum LDL-C was selected as the fourth variable of split, with an optimal cutoff of 120 mg/dL. Patients with higher LDL-C level had the higher probability of sustained virological response (83% versus 50%). Stage of fibrosis was selected as the final variable of split, with significant fibrosis (F2–4) being the predictor of lower sustained virological response probability (64% versus 32%).

Among the 6 subgroups derived by this decision tree, the subgroup of patients with 2 or more substitutions in ISDR (subgroup 1) or with a single or no substitution in ISDR but younger than 60 years of age, having the wild-type Core70 and high serum level of LDL-C (≥120 mg/dL) (subgroup 2) showed the highest probability of sustained virological response (83%).

Validation of the decision-tree model

The decision-tree model was validated using a validation dataset of 201 cases that were not included the model-building dataset. Each patient in the validation set was allocated to subgroups 1–6 using the flowchart form of the decision tree. The rates of sustained virological response were 75% for subgroup 1, 73% for subgroup 2, 65% for subgroup 3, 41% for subgroup 4, 46% for subgroup 5, and 33% for subgroup 6. The rates of sustained virological response for each subgroup of patients were correlated closely between the model building dataset and the validation dataset ($r^2 = 0.94$) (Figure 2).

The 6 subgroups were reconstructed into 3 groups according to their rate of sustained virological response: the high-probability group consisted of subgroups 1 and 2, the intermediate-probability group consisted of subgroup 3, and the low-probability group consisted of subgroups 4, 5, and 6. The rate of sustained virological response in the high-probability group was high on a consistent basis: 83% for model-building patients and 74% for validation patients. The rate of sustained virological response in the intermediate-probability group was 64% for model building patients and 65% for internal validation patients. The rate of sustained virological response in the low-probability group was low on a consistent basis: 32% for model-building patients and 36% for internal validation patients (Figure 3). Thirty percent of the patients were classified into the high-probability group and 10% of the patients were classified into intermediate-probability group, which means that about 40% of patients with higher than average probability of achieving sustained virological response were identified.

Effect of dose reductions of PEG-IFN and RBV

The possible effect of drug reductions was analyzed in the 3 groups of patients divided by decision tree (low-, intermediate-, and high-probability groups) (Figure 4). Patients were stratified according to the cumulative drug exposure with PEG-IFN and RBV: the good adherence group consisted of patients who took ≥80% planned doses of both PEG-IFN and RBV; the poor adherence group consisted of patients who took <80% of planned doses of both PEG-IFN and RBV. Even after adjustment for drug adherence, the 3 groups of patients divided by decision-tree analysis still had low, intermediate, and high probability of achieving sustained virological response, respectively, indicating that this model predicts sustained virological response independent of drug exposure.

Multivariable logistic regression analysis

Age, sex, serum levels of creatinine, ALT, GGT, LDL-C, hemoglobin, platelet count, HCV RNA titer, ISDR substitution, substitution at Core70, substitution at Core91, histological stage of fibrosis, and grade of activity were found to be associated with sustained virological response by standard univariable analysis. Multivariable analysis including these factors showed that age, sex, LDL-C levels, GGT levels, platelet count, ISDR substitution, and substitution at Core70 showed independent associations with sustained virological response (Table 2). Substitution in ISDR had the highest odds ratio, at 9.92. Fibrosis, which was selected as a significant predictor of response in the decision-tree analysis, was not found to be an independent predictor of response in standard multivariable analysis, indicating that the decision-tree analysis could identify significant predictors that would apply specifically to selected patients.