Table 9 Transcatheter arterial embolization

	HCC		ICC		Combined	
	$n = 17\,898$		$n = 736$		$n = 149$	
Not performed	9 710	(54.3%)	707	(96.1%)	113	(75.8%)
Performed	8 188	(45.7%)	29	(3.9%)	36	(24.2%)
Embolic materials	$n = 7\,850$		$n = 28$		$n = 37$	
Lipiodol	1 621	(20.6%)	8	(28.6%)	16	(43.2%)
Gelatin sponge	205	(2.6%)	1	(3.6%)	0	(0.0%)
Lipiodol + gelatin sponge	5 936	(75.6%)	18	(64.3%)	21	(56.8%)
Others	88	(1.1%)	1	(3.6%)	0	(0.0%)
Extent of embolization	$n = 7\,157$		$n = 26$		$n = 34$	
Less than one segment	2 578	(36.0%)	8	(30.8%)	6	(17.6%)
One segment to one lobe	2 896	(40.5%)	8	(30.8%)	16	(47.1%)
More than one lobe	1 252	(17.5%)	4	(15.4%)	7	(20.6%)
Whole liver	431	(6.0%)	6	(23.1%)	5	(14.7%)
Efficacy evaluation at 6 months	$n = 5\,448$		$n = 13$		$n = 24$	
CR	2 208	(40.5%)	4	(30.8%)	3	(12.5%)
PR	1 502	(27.6%)	1	(7.7%)	5	(20.8%)
SD	632	(11.6%)	3	(23.1%)	6	(25.0%)
PD	1 106	(20.3%)	5	(38.5%)	10	(41.7%)

Combined, combined hepatocellular and cholangiocarcinoma; CR, complete response; HCC, hepatocellular carcinoma; ICC, intrahepatic cholangiocarcinoma; MR, minor response; NC, no change; PD, progressive disease; PR, partial response.

LCSGJ was estimated from data collected in the surveys.

ICC and combined HCC and ICC

For ICC, cumulative survival rates were calculated for all patients and based on various background factors. For combined HCC and ICC, cumulative survival rates were calculated for all patients (Tables 14,15).

Changes in the cumulative survival rates of HCC patients

The cumulative survival rates of newly-registered HCC patients in the 5th to 18th follow-up surveys (1978–2005) whose final prognosis was defined as survival or death (excluding cases of unknown outcome) divided into three groups (1978–1985, 1986–1995 and 1996–2005) were also calculated (Fig. 1). The 3- and 5-year cumulative survival rates were 15.7% and 9.5% in patients between 1978 and 1985 ($n = 7852$), 42.1% and 26.8% between 1986 and 1995 ($n = 51\,719$), and 56.6% and 39.3% between 1996 and 2005 ($n = 88\,590$), respectively.

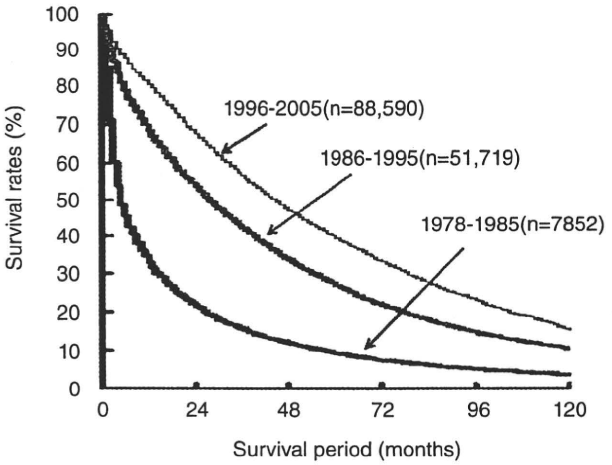

Figure 1 Cumulative survival rates of newly-registered patients in the 5th to 18th follow-up surveys (1978–2005) divided into three groups (1978–1985, 1986–1995 and 1996–2005) are shown. The 3- and 5-year cumulative survival rates were 15.7%, 9.5% in patients between 1978 and 1985 ($n = 7852$), 42.1% and 26.8% between 1986 and 1995 ($n = 51\,719$), and 56.6% and 39.3% between 1996 and 2005 ($n = 88\,590$), respectively.

Table 10 Microscopic pathological findings of surgical or biopsy specimens

	HCC		ICC		Combined	
Capsule formation	$n = 5221$		$n = 406$		$n = 84$	
Fc$^-$	1293	(24.8%)	386	(95.1%)	54	(64.3%)
Fc$^+$	3928	(75.2%)	20	(4.9%)	30	(35.7%)
Capsule infiltration	$n = 3850$		$n = 16$		$n = 30$	
Fc-inf$^-$	1264	(32.8%)	8	(50.0%)	8	(26.7%)
Fc-inf$^+$	2586	(67.2%)	8	(50.0%)	22	(73.3%)
Septum formation	$n = 4983$		$n = 372$		$n = 83$	
Sf$^-$	1930	(38.7%)	348	(93.5%)	41	(49.4%)
Sf$^+$	3053	(61.3%)	24	(6.5%)	42	(50.6%)
Serosal invasion	$n = 4959$		$n = 409$		$n = 82$	
S0	4267	(86.0%)	267	(65.3%)	61	(74.4%)
S1	537	(10.8%)	96	(23.5%)	15	(18.3%)
S2	84	(1.7%)	44	(10.8%)	5	(6.1%)
S3	71	(1.4%)	2	(0.5%)	1	(1.2%)
Lymph node metastasis	$n = 3984$		$n = 427$		$n = 70$	
Absent	3938	(98.8%)	257	(60.2%)	57	(81.4%)
Present	46	(1.2%)	170	(39.8%)	13	(18.6%)
Portal vein invasion	$n = 5368$		$n = 430$		$n = 87$	
vp0	3971	(74.0%)	223	(51.9%)	41	(47.1%)
Vp1	1019	(19.0%)	137	(31.9%)	33	(37.9%)
Vp2	167	(3.1%)	37	(8.6%)	6	(6.9%)
Vp3	138	(2.6%)	31	(7.2%)	7	(8.0%)
Vp4	73	(1.4%)	2	(0.5%)	0	(0.0%)
Hepatic vein invasion	$n = 5320$		$n = 423$		$n = 84$	
Vv0	4714	(88.6%)	304	(71.9%)	61	(72.6%)
Vv1	499	(9.4%)	85	(20.1%)	23	(27.4%)
Vv2	77	(1.4%)	24	(5.7%)	0	(0.0%)
Vv3	30	(0.6%)	10	(2.4%)	0	(0.0%)
Hepatic arterial invasion	$n = 5160$		$n = 402$		$n = 82$	
Va0	5103	(98.9%)	377	(93.8%)	79	(96.3%)
Va1	54	(1.0%)	18	(4.5%)	2	(2.4%)
Va2	2	(0.0%)	3	(0.7%)	1	(1.2%)
Va3	1	(0.0%)	4	(1.0%)	0	(0.0%)
Bile duct invasion	$n = 5279$		$n = 403$		$n = 87$	
B0	5095	(96.5%)	184	(45.7%)	66	(75.9%)
B1	108	(2.0%)	91	(22.6%)	15	(17.2%)
B2	37	(0.7%)	50	(12.4%)	3	(3.4%)
B3	21	(0.4%)	61	(15.1%)	1	(1.1%)
B4	18	(0.3%)	17	(4.2%)	2	(2.3%)
Intrahepatic metastasis	$n = 5206$		$n = 430$		$n = 86$	
Im0	4147	(79.7%)	322	(74.9%)	52	(60.5%)
Ims	238	(4.6%)	17	(4.0%)	5	(5.8%)
Im1	384	(7.4%)	39	(9.1%)	11	(12.8%)
Im2	299	(5.7%)	34	(7.9%)	10	(11.6%)
Im3	138	(2.7%)	18	(4.2%)	8	(9.3%)
Surgical margin	$n = 5104$		$n = 434$		$n = 84$	
Presence of cancer invasion	408	(8.1%)	80	(18.4%)	13	(15.5%)
Absence of cancer invasion	4696	(91.9%)	354	(81.6%)	71	(84.5%)
Non-cancerous portion	$n = 5395$		$n = 414$		$n = 84$	
Normal liver	349	(6.5%)	269	(65.0%)	9	(10.7%)
Chronic hepatitis or liver fibrosis	2587	(48.0%)	101	(24.4%)	46	(54.8%)
Liver cirrhosis	2459	(45.6%)	44	(10.6%)	29	(34.5%)
Liver fibrosis	$n = 3153$		$n = 169$		$n = 49$	
F0 (normal)	184	(5.8%)	82	(48.5%)	5	(10.2%)
F1	429	(13.6%)	39	(23.1%)	3	(6.1%)
F2	532	(16.9%)	14	(8.3%)	12	(24.5%)
F3	578	(18.3%)	13	(7.7%)	12	(24.5%)
F4 (liver cirrhosis)	1430	(45.4%)	21	(12.4%)	17	(34.7%)

B0–B4, described in Tables 5 and 7; combined, combined hepatocellular and cholangiocarcinoma; Fc, Fc-inf, described in Table 7; F1, fibrosis expansion of portal tract; F2, bridging fibrosis formation; F3, bridging fibrosis formation accompanying lobular distortion; HCC, hepatocellular carcinoma; ICC, intrahepatic cholangiocarcinoma; Im0–Im3, described in Table 7; Sf, S0–S3 described in Table 7; Va0–Va3, described in Table 7; Vp0–Vp4, Vv0–Vv3, described in Tables 5 and 7.

Table 11 Cumulative survival rates (%) of HCC patients treated with hepatic resection (1994–2005)

		n	Year									
			1	2	3	4	5	6	7	8	9	10
All cases		25 066	88.2%	78.4%	69.5%	61.7%	54.2%	48.1%	42.0%	36.9%	32.5%	29.0%
Tumor size (cm)	≤2	4 363	95.8%	91.1%	85.4%	78.2%	69.4%	61.7%	53.4%	46.5%	40.5%	35.5%
	2–5	12 801	91.9%	82.9%	73.2%	65.0%	56.8%	50.2%	43.9%	38.8%	34.2%	30.6%
	5–10	4 802	82.3%	68.7%	58.5%	50.2%	44.0%	39.1%	34.0%	29.8%	26.0%	23.6%
	>10	2 044	66.5%	50.6%	42.5%	36.7%	32.1%	29.5%	25.9%	22.6%	20.3%	18.5%
Tumor number	1	17 531	91.0%	82.9%	74.8%	67.7%	60.2%	54.0%	47.5%	42.1%	37.5%	33.2%
	2	3 692	87.3%	75.3%	64.8%	55.9%	48.0%	40.3%	34.8%	28.5%	24.6%	22.7%
	≥3	3 010	75.7%	59.6%	48.1%	38.4%	30.6%	26.3%	22.0%	19.3%	15.3%	13.7%
Portal vein invasion	Vp0	20 195	92.2%	83.7%	74.9%	67.0%	59.0%	52.4%	45.5%	40.1%	35.3%	31.3%
	Vp1	1 978	79.3%	64.9%	54.2%	45.7%	39.1%	34.3%	31.9%	28.1%	24.2%	22.9%
	Vp2	820	61.0%	45.4%	33.6%	27.6%	23.3%	22.8%	20.6%	17.0%	16.0%	16.0%
	Vp3 or Vp4	1 021	52.1%	33.6%	26.4%	22.4%	18.3%	16.6%	14.8%	13.1%	10.5%	8.4%
Non-cancerous portion	Normal liver	1 801	86.2%	76.2%	68.9%	63.6%	59.1%	55.7%	51.1%	46.9%	43.4%	37.6%
	Chronic hepatitis/ liver fibrosis	9 581	90.4%	81.5%	73.4%	67.0%	60.8%	55.8%	50.2%	45.6%	41.7%	39.0%
	Liver cirrhosis	10 401	87.3%	77.0%	67.3%	58.3%	49.1%	42.1%	35.1%	30.2%	25.4%	22.1%
Liver damage classification by LCSGJ	A	16 963	90.0%	81.5%	73.3%	66.0%	59.0%	52.9%	46.3%	41.5%	36.7%	33.2%
	B	6 478	85.6%	73.8%	63.6%	54.8%	45.3%	39.2%	33.8%	28.6%	25.1%	21.3%
	C	454	73.4%	56.0%	44.9%	39.8%	35.0%	32.1%	30.9%	22.9%	21.7%	21.7%
TNM Stage by LCSGJ	I	2 846	96.9%	93.6%	88.7%	81.8%	73.0%	66.1%	57.6%	51.3%	45.4%	38.1%
	II	12 458	92.7%	84.1%	75.3%	67.4%	59.7%	53.4%	46.1%	40.4%	35.9%	32.5%
	III	4 223	82.2%	68.1%	56.1%	47.2%	39.5%	34.1%	30.6%	26.9%	23.6%	21.4%
	IV A	1 398	60.3%	42.4%	31.9%	25.9%	21.4%	19.7%	17.8%	15.3%	12.5%	11.9%
	IV B	253	53.1%	33.6%	24.2%	21.7%	16.5%	14.1%	14.1%	14.1%	14.1%	14.1%

HCC, hepatocellular carcinoma; LCSGJ, Liver Cancer Study Group of Japan; TNM, Tumor–Node–Metastasis; Vp0–V94, described in Tables 5 and 7.

Table 12 Cumulative survival rates (%) of HCC patients treated with local ablation therapy (1994–2005)

		n	Year									
			1	2	3	4	5	6	7	8	9	10
All cases		27 150	92.8%	81.4%	68.6%	56.5%	45.6%	37.1%	29.8%	23.9%	19.5%	15.7%
Liver damage classification by LCSGJ	A	14 370	95.5%	87.2%	76.3%	65.5%	54.2%	44.4%	36.6%	30.3%	25.0%	19.9%
	B	9 751	92.4%	78.5%	63.4%	50.0%	38.7%	31.0%	24.2%	18.0%	14.7%	12.4%
	C	1 757	77.2%	56.2%	41.2%	28.1%	21.6%	16.9%	12.3%	9.4%	7.1%	5.4%
Tumor number	1	16 883	94.2%	84.5%	73.2%	62.3%	51.9%	42.8%	35.1%	28.8%	23.8%	19.4%
	2	5 638	92.4%	79.8%	65.9%	51.8%	39.6%	32.8%	24.4%	18.7%	15.5%	11.7%
	3	2 307	91.6%	76.8%	60.9%	46.3%	35.0%	25.8%	20.6%	15.7%	11.5%	10.9%
	4	812	88.5%	72.6%	55.3%	41.1%	30.8%	21.4%	17.7%	13.1%	6.9%	3.4%
	≥5	1 079	82.9%	62.1%	44.0%	33.0%	23.4%	20.4%	13.9%	12.2%	9.1%	7.2%
Tumor size (cm)	≤1	1 792	96.6%	90.6%	81.6%	72.1%	60.1%	49.8%	44.6%	38.4%	31.1%	25.7%
	1–2	12 253	95.2%	86.4%	75.1%	63.7%	52.7%	43.0%	34.0%	27.3%	22.0%	18.1%
	2–3	7 714	93.0%	79.7%	64.8%	51.9%	40.0%	32.1%	25.5%	20.1%	16.4%	13.4%
	3–5	3 257	88.2%	71.0%	55.7%	41.8%	32.4%	26.1%	20.8%	17.6%	15.3%	8.7%
	>5	809	76.9%	58.9%	43.6%	33.7%	25.5%	21.0%	15.6%	10.6%	9.1%	–

HCC, hepatocellular carcinoma; LCSGJ, Liver Cancer Study Group of Japan.

Newly-registered patients were increasing and their survival rates were improving.

DISCUSSION

PRIMARY LIVER CANCER is the fourth leading cause of cancer death in Japanese people, following tracheal–bronchial–lung, gastric and colorectal cancers; more than 34 000 people die annually due to liver cancer. In the 18th Nationwide Follow-Up Survey of Primary Liver Cancer, approximately 30% of patients with primary liver cancer were newly registered. Compared with the 17th follow-up survey,[11] this follow-up survey in HCC indicated an increase in elder patients and women, a decrease in patients positive for hepatitis B surface antigen and hepatitis C virus antibody, and a decrease in tumor size at the clinical diagnosis. In the local ablation therapy, ratio of radio frequency ablation therapy was increasing. Advance in diagnostic and therapeutic modalities were considered to have contributed to an improvement in survival of patients with HCC between 1978 and 2005.

We hope that the results of this follow-up survey will contribute to research and improved medical practice for primary liver cancer.

ACKNOWLEDGMENTS

WE WOULD LIKE to express our sincere gratitude to the doctors of the 544 medical institutions that participated in this follow-up survey, and to Mrs T. Idutsu and M. Ogawa for data compilation.

REFERENCES

1 Okuda K, The Liver Cancer Study Group of Japan. Primary liver cancers in Japan. *Cancer* 1980; **45**: 2663–9.
2 The Liver Cancer Study Group of Japan. Primary liver cancer in Japan. *Cancer* 1984; **54**: 1747–55.
3 The Liver Cancer Study Group of Japan. Primary liver cancer in Japan-Sixth report. *Cancer* 1987; **60**: 1400–11.
4 The Liver Cancer Study Group of Japan. Primary liver cancer in Japan. *Ann Surg* 1990; **211**: 277–87.
5 Primary Liver Cancer in Japan. Tobe T *et al.* Springer-Verlag Tokyo, Berlin, Heidelberg, New York, London, Paris, Hong Kong, Barcelona 1992.
6 The Liver Cancer Study Group of Japan. Predictive factors for long term prognosis after partial hepatectomy for patients with hepatocellular carcinoma in Japan. *Cancer* 1994; **74**: 2772–80.
7 Arii S, Yamaoka Y, Futagawa S *et al.* Results of surgical and nonsurgical treatment for small-sized hepatocellular carci-

Table 13 Cumulative survival rates (%) of HCC patients treated with transcatheter arterial embolization (1994–2005)

		n	Year									
			1	2	3	4	5	6	7	8	9	10
All cases		3955	51.7%	35.1%	28.5%	23.7%	20.3%	18.1%	16.7%	14.5%	12.5%	12.5%
Liver damage classification by LCSGJ	A	1658	72.0%	53.1%	43.9%	36.2%	31.3%	27.6%	26.1%	23.7%	21.2%	21.2%
	B	2294	36.3%	21.6%	17.1%	14.4%	12.2%	11.0%	9.6%	5.8%	0.0%	–
	C	137	88.3%	81.3%	75.2%	67.9%	62.8%	59.8%	59.8%	54.8%	54.8%	54.8%
Tumor number	1	738	77.8%	58.9%	49.4%	40.1%	32.3%	26.5%	25.3%	23.6%	18.2%	18.2%
	2	547	63.7%	43.4%	33.4%	29.0%	26.7%	24.9%	23.6%	20.4%	18.1%	18.1%
	3	129	55.3%	32.8%	28.6%	22.2%	19.0%	14.2%	14.2%	14.2%	14.2%	14.2%
	4	1272	76.3%	58.8%	49.4%	41.2%	36.3%	32.2%	30.8%	28.4%	24.8%	24.8%
	≥5	102	75.7%	49.4%	36.5%	31.3%	27.8%	27.8%	22.2%	22.2%	22.2%	22.2%

HCC, hepatocellular carcinoma; LCSGJ, Liver Cancer Study Group of Japan.

Table 14 Cumulative survival rates (%) of ICC patients (1994–2005)

			n	Year									
				1	2	3	4	5	6	7	8	9	10
All cases			3955	51.7%	35.1%	28.5%	23.7%	20.3%	18.1%	16.7%	14.5%	12.5%	12.5%
Hepatic resection		Performed	1658	72.0%	53.1%	43.9%	36.2%	31.3%	27.6%	26.1%	23.7%	21.2%	21.2%
		Not performed	2294	36.3%	21.6%	17.1%	14.4%	12.2%	11.0%	9.6%	5.8%	0.0%	–
Cases of hepatic resection	Tumor size (cm)	≤2	137	88.3%	81.3%	75.2%	67.9%	62.8%	59.8%	59.8%	54.8%	54.8%	54.8%
		2–5	738	77.8%	58.9%	49.4%	40.1%	32.3%	26.5%	25.3%	23.6%	18.2%	18.2%
		5–10	547	63.7%	43.4%	33.4%	29.0%	26.7%	24.9%	23.6%	20.4%	18.1%	18.1%
		>10	129	55.3%	32.8%	28.6%	22.2%	19.0%	14.2%	14.2%	14.2%	14.2%	14.2%
	Tumor number	1	1272	76.3%	58.8%	49.4%	41.2%	36.3%	32.2%	30.8%	28.4%	24.8%	24.8%
		2	102	75.7%	49.4%	36.5%	31.3%	27.8%	27.8%	22.2%	22.2%	22.2%	22.2%
		≥3	186	42.2%	19.5%	16.6%	12.3%	6.3%	4.2%	4.2%	2.1%	2.1%	2.1%
	Residual tumor	Absent	784	77.7%	59.3%	50.6%	43.1%	37.6%	35.6%	33.6%	30.2%	26.5%	26.5%
		Present	608	64.4%	41.4%	31.3%	22.1%	20.6%	20.6%	10.3%	10.3%	10.3%	–
	Lymph node metastasis	Absent	1046	80.6%	64.6%	54.5%	45.3%	39.9%	36.2%	33.8%	30.2%	28.8%	28.8%
		Present	497	55.9%	29.8%	22.8%	17.9%	15.3%	10.7%	10.7%	10.7%	8.0%	8.0%

ICC, intrahepatic cholangiocarcinoma.

Table 15 Cumulative survival rates (%) of combined HCC and ICC (1994–2005)

		n	Year									
			1	2	3	4	5	6	7	8	9	10
All cases		653	58.6%	40.5%	29.7%	23.4%	19.8%	17.8%	15.7%	14.5%	12.7%	12.7%
Hepatic resection	Performed	354	70.7%	50.5%	40.7%	31.0%	28.2%	26.1%	21.9%	20.0%	20.0%	20.0%
	Not performed	299	44.2%	28.8%	16.9%	14.2%	10.6%	8.9%	8.9%	8.9%	0.0%	–

HCC, hepatocellular carcinoma; ICC, intrahepatic cholangiocarcinoma.

nomas: a retrospective and nationwide survey in Japan. The liver cancer study group of Japan. *Hepatology* 2000; **32**: 1224–9.
8 Ikai I, Itai Y, Okita K *et al.* Report of the 15th follow-up survey of primary liver cancer. *Hepatol Res* 2004; **28**: 21–9.
9 Ikai I, Arii S, Kojiro M *et al.* Re-evaluation of prognostic factors for survival after liver resection in patients with hepatocellular carcinoma in a Japanese nationwide survey. *Cancer* 2004; **101**: 796–802.
10 Ikai I, Arii S, Ichida T *et al.* Report of the 16th follow-up survey of primary liver cancer. *Hepatol Res* 2005; **32**: 163–72.
11 Ikai I, Arii S, Okazaki M *et al.* Report of the 17th nationwide follow-up survey of primary liver cancer. in Japan. *Hepatol Res* 2007; **37**: 676–91.
12 Takayasu K, Arii S, Ikai I *et al.* Prospective cohort study of transarterial chemoembolization for unresectable hepatocellular carcinoma in 8510 patients. *Gastroenterology* 2006; **131**: 461–9.
13 Ikai I, Takayasu K, Omata M *et al.* A modified Japan integrated stage score for prognostic assessment in patients with hepatocellular carcinoma. *J Gastroenterol* 2006; **41**: 884–92.
14 Minagawa M, Ikai I, Matsuyama Y *et al.* Staging of Hepatocellular Carcinoma: assessment of the Japanese TNM and AJCC/UICC TNM systems in a cohort of 13 772 patients in Japan. *Ann Surg* 2007; **245**: 909–22.
15 Eguchi S, Kanematsu T, Arii S *et al.* Comparison of the outcomes between an anatomical subsegmentectomy and a non-anatomical minor hepatectomy for single hepatocellular carcinomas based on a Japanese nationwide survey. *Surgery* 2008; **143**: 469–75.
16 Hasegawa K, Makuuchi M, Takayama T *et al.* Surgical resection vs. percutaneous ablation for hepatocellular carcinoma: a preliminary report of the Japanese nationwide survey. *J Hepatol* 2008; **49**: 589–94.
17 Takayasu K, Arii S, Ikai I *et al.* Liver Cancer Study Group of Japan. Overall survival after transarterial lipiodol infusion chemotherapy with or without embolization for unresectable hepatocellular carcinoma: propensity score analysis. *AJR Am J Roentgenol* 2010; **194**: 830–7.
18 Liver Cancer Study Group of Japan. *General Rules for the Clinical and Pathological Study of Primary Liver Cancer, Second English Edition.* Tokyo: Kanehara & Co., Ltd., 2003.
19 Kudo M, Kubo S, Takayasu K *et al.* Response evaluation criteria in cancer of the liver (RECICL) proposed by the liver cancer study group of Japan (2009 revised version). *Hepatol Res* 2010; **40**: 686–62.

Research Article

Statistics in Medicine

Received 21 December 2009, Accepted 6 May 2010 Published online 15 June 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3987

A comparison of the results of intent-to-treat, per-protocol, and g-estimation in the presence of non-random treatment changes in a time-to-event non-inferiority trial

Yutaka Matsuyama*†

While intent-to-treat (ITT) analysis is widely accepted for superiority trials, there remains debate about its role in non-inferiority trials. It has often been said that ITT analysis tends to be anti-conservative in demonstrating non-inferiority, suggesting that per-protocol (PP) analysis may be preferable for non-inferiority trials, despite the inherent bias of such analyses. We propose using randomization-based g-estimation analyses that more effectively preserve the integrity of randomization than do the more widely used PP analyses. Simulation studies were conducted to investigate the impacts of different types of treatment changes on the conservatism or anti-conservatism of analyses using the ITT, PP, and g-estimation methods in a time-to-event outcome. The ITT results were anti-conservative for all simulations. Anti-conservativeness increased with the percentage of treatment change and was more pronounced for outcome-dependent treatment changes. PP analysis, in which treatment-switching cases were censored at the time of treatment change, maintained type I error near the nominal level for independent treatment changes, whereas for outcome-dependent cases, PP analysis was either conservative or anti-conservative depending on the mechanism underlying the percentage of treatment changes. G-estimation analysis maintained type I error near the nominal level even for outcome-dependent treatment changes, although information on unmeasured covariates is not used in the analysis. Thus, randomization-based g-estimation analyses should be used to supplement the more conventional ITT and PP analyses, especially for non-inferiority trials. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: g-estimation; intent-to-treat; non-compliance; non-inferiority trials; per-protocol; randomization-based analysis

1. Introduction

Most randomized clinical trials are designed to demonstrate the superiority of a new treatment over a standard treatment or placebo. An increasing number of trials, however, are focused on showing that a new treatment is not worse than a standard one (active control) by more than a pre-specified margin. One reason for conducting such non-inferiority trials is that it has become increasingly difficult to demonstrate superiority over an active control in clinical trials for certain diseases. Another reason is that a new treatment with comparable efficacy might have other advantages, such as fewer side effects, lower cost, greater convenience, or higher quality of life, over the active control. Approaches for the design, conduct, and analysis of non-inferiority trials have been discussed in several papers [1–6]. However, certain issues that arise in non-inferiority trials require further investigation.

One such critical issue is the effect of non-compliance on non-inferiority trials. Although it is important to minimize protocol deviations, such as violations of the entry criteria and non-compliance with the randomized treatments, most clinical trials are not ideal, and patients often fail to adhere to their assigned treatment and switch to another trial treatment or a non-trial treatment. When non-compliance occurs, the data are most commonly analyzed using intent -to-treat (ITT) and per-protocol (PP) approaches. In ITT analysis, patients are analyzed according to their assigned treatment, regardless of whether they actually comply with the treatment. In PP analysis, only patients who completely adhere to their treatments are included in the analysis.

Department of Biostatistics, School of Public Health, The University of Tokyo, Tokyo 113-0033, Japan
**Correspondence to: Yutaka Matsuyama, Department of Biostatistics, School of Public Health, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.*
†E-mail: matuyama@epistat.m.u-tokyo.ac.jp

Contract/grant sponsor: Ministry of Health, Labour and Welfare, Japan

For superiority trials, ITT analysis is well accepted for the primary analysis, because it avoids the overly optimistic estimates of treatment effects that can result from PP analysis and generally produces estimates that are more conservative. It also addresses a pragmatic question, that is, what is the benefit of allocating patients to a new treatment compared with allocating them to a standard one? However, there is less agreement on its role in non-inferiority trials. It is generally perceived that ITT analysis is anti-conservative in non-inferiority trials; i.e. non-inferiority is more easily demonstrated due to dilution of treatment effects resulting from non-compliance [3, 7, 8]. The International Conference on Harmonization (ICH) E9 document [9] states that analysis of the ITT population is 'generally not conservative and its role should be considered very carefully' in the context of non-inferiority trials. This statement has been interpreted to indicate that PP analysis may be preferable for non-inferiority trials. However, when there are treatment-related drop-outs, the two treatments (test treatment and active control) may appear similar in a PP analysis, because both groups of patients remaining in the study would most likely be responders [10]. Therefore, the current thinking of regulatory agencies is that the study objective should be achieved for both ITT and PP populations, especially for non-inferiority trials [3]. The Committee on Proprietary Medical Products Points-to-Consider [11] specifically states that '...similar conclusions from both the ITT and PP are required in a non-inferiority trial'.

There are several papers that have investigated the impacts of non-compliance on ITT and PP analyses for equivalence and non-inferiority trials [7, 10, 12–15]. Rohmel [12] concluded that in the presence of non-compliance, an ITT analysis results in a higher rate of erroneous conclusions of equivalence. On the other hand, other authors [10, 13–15] have concluded that the bias can be in either direction for both ITT and PP analyses, depending on the mechanism underlying the probability of non-compliance, alternative treatment received by the non-compliers, event probabilities, and other factors.

It is clear that both ITT and PP analyses are problematic in the presence of non-random non-compliance, because of its effects on estimates of treatment effects, type I error, and power of the study. In the context of equivalence trials, Robins [16] provided an overview of methods that can be used to adjust for non-compliance. Robins and colleagues recommended using methods such as inverse probability of censoring-weighted estimators, inverse probability of treatment-weighted estimators of marginal structural models, and g-estimators of structural nested models. In a new drug application setting, it is generally preferable to use statistical methods that are less sensitive to the underlying assumptions of their approaches. All of the above approaches are, in general, sensitive to their underlying assumptions, because some assumptions must be made to estimate the actual benefit of intervention, i.e. causal effects. However, for non-inferiority trials, obtaining reliable compliance data and implementing methods that adjust for the effects of non-compliance in the analysis should also be considered to supplement the more conventional ITT and PP analyses. In this paper, although we assume the structural model for the causal parameter, we have chosen to avoid all assumptions about either observed or unobserved factors that influence an individual's decision to comply, while comparing outcomes based only on the treatment groups randomized by design, i.e. a randomization-based analysis. Robins and Tsiatis [17] and Mark and Robins [18] introduced a randomization-based analysis for time-to-event data based on rank preserving structural failure time (RPSFT) models. This analysis more effectively preserves the integrity of randomization than does the more widely used PP analysis.

The aim of this paper is to investigate the impacts of different types of treatment changes on ITT, PP, and g-estimation analyses for non-inferiority trials with a time-to-event outcome. In Section 2, the randomization-based g-estimation analysis method is briefly reviewed and the simulation studies are described. Section 3 presents the simulation results. Finally, in Section 4, we provide a brief discussion and conclusions.

2. Methods

2.1. G-estimation method based on the randomized analysis

Consider a randomized clinical trial for non-inferiority in which two groups (test and standard treatments) are compared with respect to a time-to-event outcome, but each patient $i (i = 1, \ldots, N)$ may fail to comply with the assigned treatment $R_i (R_i = 1$ if allocated to the test, $R_i = 0$ if allocated to the standard) and cross over to the other treatment at time t ($t > 0$; time zero is the randomization time and the start of treatment).

Suppose we have repeated measures on the actual treatment status at time t, $A_i(t)$ ($A_i(t) = 1$ if the test, $A_i(t) = 0$ if the standard). For patient i, we define U_i as the potential failure time patient i would have if the patient had received standard treatment all the time throughout the study period. Robins and coworkers [17, 18] introduced an RPSFT model, in which each patient's baseline failure time U_i is related to his/her observable data:

$$U_i = \int_0^{T_i} \exp[-\psi_0 A_i(t)]\,\mathrm{d}t, \tag{1}$$

where T_i is the observed failure time for patient i and ψ_0 is an unknown parameter representing the causal effect of the test treatment versus the standard treatment. Our notation for the potential outcomes implicitly assumes Rubin's stable unit treatment value assumption (SUTVA), which implies that the potential outcomes for patient i do not depend on the treatment received by any other patient [19]. We also assume that the potential outcomes satisfy the consistency assumption [20, 21], which states that an individual's potential outcome under his/her observed treatment history is precisely his/her observed outcome regardless of means (route, condition, etc.) of taking the treatment. Note also that U_i is defined for each individual at the time of randomization and is treated as a baseline fixed characteristic, like sex or age.

The RPSFT model (1) can be represented as an accelerated failure time model with a time-varying treatment, with $A_i(t)=1$ if patient i is on the test treatment and $A_i(t)=0$ if patient i is on the standard treatment at time t. When $A_i(t)\equiv 0$ for all t, then (1) gives $T_i=U_i$ as expected, whereas when $A_i(t)\equiv 1$ for all t, (1) gives $T_i=U_i\exp(\psi_0)$. Therefore, the ratio $\exp(\psi_0)$ is the expansion (or contraction) in failure time when comparing continuous test treatment versus continuous standard treatment. Negative values of ψ_0 indicate that a patient's failure time with the test treatment is shorter than the failure time with the standard treatment. In the analysis of time-to-event data, the proportional hazards model is widely used to estimate the causal hazard ratio, i.e. the hazard had all patients been continuously treated with the test treatment divided by the hazard had all patients been continuously treated with the standard treatment. When the underlying distribution of U_i is a Weibull distribution with parameters ρ and k, i.e. $\Pr[U_i>t]=\exp[-(\rho t)^k]$, there is a one-to-one correspondence between ψ_0 in model (1) and the causal hazard ratio, i.e. the causal hazard ratio equals $\exp(-k\psi_0)$ [22].

Randomization guarantees that any variable unmeasured as well as measured at baseline will on average be balanced with respect to the treatment assignment. Specifically, U_i is independent of R_i, because U_i is a fixed characteristic of the individual. Thus, a procedure for estimating ψ_0 is as follows. We define a random variable $U_i(\psi)$ as equal to the right-hand side of (1) for a given ψ. We also define $Z(\psi)$ as a test statistic comparing the distribution of $U_i(\psi)$ for the two randomized groups, using the log-rank test. In this log-rank statistic, $U_i(\psi)$ for each subject are treated as though they are the observed failure time, and its numerator is a sum of the observed R_i minus the expected value of R_i when R_i is randomly chosen from the risk set $Y_i(\psi)=\{j: U_j(\psi)\geqslant U_i(\psi)\}$,

$$\sum_{i=1}^{N}\left(R_i-\frac{\sum_{j\in Y_i(\psi)}R_j}{n_i(\psi)}\right),$$

where $n_i(\psi)$ is the number of subject in the risk set $Y_i(\psi)$. The variance of this statistic can be consistently estimated by the usual formulas for the variance of a log-rank test. The point estimate of ψ_0 (the g-estimate) is the value at which $Z(\psi)=0$, and can be found by a search over a grid. The $100(1-\alpha)$ per cent confidence interval for ψ_0 is the range of values over which $|Z(\psi)|<z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution.

To test the hypothesis that $\psi_0=\psi$ (g-test of $\psi_0=\psi$), the p-value can easily be calculated by considering the association of the assignment R_i with $U_i(\psi)$. This test is a valid, randomization-based (ITT) test of the hypothesis. In particular, the g-test of the null for no treatment effect is simply an ITT test of the effect of treatment comparing $R_i=1$ and $R_i=0$. This equivalence demonstrates that g-estimation maintains the original randomized group, whereas PP analysis does not.

One complication arises from the fact that due to censoring $U_i(\psi)$ cannot always be computed from the observed data. Here, we assume that reaching time C without death is the only source of censoring, where C is the time from randomization to the fixed end of the study. If T_i is a censored time, then $U_i(\psi)$ is censored at

$$D_i(\psi)=\int_0^{C_i}\exp[-\psi A_i(t)]\,dt,$$

where C_i is the potential censoring time for each patient. Although C_i is known for uncensored as well as censored subjects, $D_i(\psi)$ is a function of $A_i(t)$ and may depend on the underlying prognosis. Therefore, even when censoring on the T-scale is non-informative, i.e. an administrative censoring, censoring on the U-scale is likely to be informative if $\psi_0\neq 0$ and there is non-random non-compliance. Thus, we cannot replace T_i by $X_i=\min(T_i, C_i)$ to calculate the pseudo-baseline event time.

To resolve this issue, Robins and Tsiatis [17] defined a new censoring time $C_i(\psi)=C_i$ if $\psi\leqslant 0$ and $C_i(\psi)=C_i\exp(-\psi)$ if $\psi>0$, based on the direction of the treatment effect. For given ψ, let $X_i(\psi)=\min[C_i(\psi), U_i(\psi)]$ and $\Delta_i(\psi)=I[U_i(\psi)>C_i(\psi)]$ be the new failure time and censoring indicator, respectively. $X_i(\psi)$ is observable, because $T_i\geqslant C_i$ implies $U_i(\psi)>C_i(\psi)$. Because any function of $\{U_i(\psi), C_i\}$ is independent of the random treatment assignment R_i, we have $\{U_i(\psi_0), \Delta_i(\psi_0)\}\coprod R_i$. Thus, treating the pair $\{X_i(\psi), \Delta_i(\psi)\}$ as the failure time and the censoring indicator, we use the log-rank test as a measure of equality for the distributions of the baseline failure time for the two randomized

groups. Hernan *et al.* [23] and Greenland *et al.* [24] provide introductory reviews for the g-estimation analysis based on the RPSFT model (1).

2.2. *Assessment of non-inferiority*

Consider a non-inferiority trial in which the test (T) and standard (S) treatment are compared with respect to overall survival. The null and alternative hypotheses for assessing non-inferiority are expressed as

$$H_0: \frac{\lambda_T}{\lambda_S} \geqslant \delta, \quad H_1: \frac{\lambda_T}{\lambda_S} < \delta,$$

where λ_S and λ_T are the hazards of death for the two groups, and $\delta(>1)$ is the non-inferiority margin. The null hypothesis states that the test treatment is inferior to the standard treatment by δ or more. The alternative hypothesis implies that a hazard ratio of less than δ is considered to be clinically acceptable and that the test treatment is at least as good as the standard one.

To assess whether the null hypothesis is rejected, we can perform a one-sided hypothesis test at an α level of significance [3]. Equivalently, we can compute a $100(1-2\alpha)$ per cent two-sided confidence interval for the hazard ratio. If the upper limit of the confidence interval is less than the pre-specified non-inferiority margin δ, then with $100(1-2\alpha)$ per cent confidence, we can state that the standard treatment is more efficacious than the test treatment by no more than δ, thus allowing us to claim non-inferiority of the test treatment compared with the standard treatment at an α level of significance.

2.3. *Simulation framework*

We assumed exponential failure times, set the overall survival at 5 years for the standard treatment at 0.75 (i.e. hazard for death, $\lambda_S, = -\log(0.75)/5$), and set the non-inferiority margin at $\delta = 1.24$, which is nearly equal to $\log(0.70)/\log(0.75)$. For a 6-year follow-up period ($C=6$) with a one-sided α level of 0.025 and a sample size of $n=1400$ per group, the study had 80 per cent power to show non-inferiority [4], if all patients complied fully with their allocated treatments. We assumed that the time from randomization to the fixed end of the study without death was the only source of censoring. To evaluate type I error rates for the ITT, PP, and g-estimation analyses in the presence of non-random treatment changes, we simulated data such that T was inferior to S with a true hazard ratio of 1.24, which was the pre-specified non-inferiority margin. Note that in a non-inferiority trial, the type I error is the probability of erroneously concluding non-inferiority when a true clinically important treatment effect exists.

We simulated data from two treatment groups, coded as $R=0$ (standard treatment) and $R=1$ (test treatment). An equal sample size of 1400 for each group was randomly generated (total sample size = 2800). For each subject i ($i=1,\ldots,2800$), a baseline covariate L_{i1} was generated from a standard normal distribution. The potential baseline failure time U_i was generated using the following linear transformation model [25, 26]:

$$\log(U_i) = -L_i^{\mathrm{T}}\beta + \varepsilon_i, \tag{2}$$

where $L_i = (1, L_{i1})^{\mathrm{T}}$, $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$, and ε_i follows the extreme value distribution with distribution function $\Pr(\varepsilon < y) = 1 - \exp[-\exp(y)]$. The draws from model (2) were exponentially distributed with a rate parameter of $\exp(L_i^{\mathrm{T}}\beta)$. The parameter β_0 was set to $\log(0.06)$, which corresponds to about 30 per cent overall death rate for the whole study period of 6-year follow-up in the standard treatment group.

The potential time from randomization until treatment change, which was denoted by D_i, was generated using the same model as (2) with parameter $\gamma = (\gamma_0, \gamma_1)^{\mathrm{T}}$. The parameter γ_0 was set to -25, $\log(0.034)$, $\log(0.075)$ or $\log(0.126)$, corresponding to about 0, 15, 30, and 45 per cent treatment change in each group, respectively. These numbers are the percentage of subjects that change their treatment once during the 5-year treatment period. We set $\beta_1 = \gamma_1 = 1$ for dependence of the potential baseline failure time on the alternative treatment (outcome-dependent treatment change) and $\beta_1 = \gamma_1 = 0$ for independence (independent treatment change).

The failure time T_i was calculated using model (1). With $\psi_0 = -\log(\delta)$, $T_i = D_i^* + [U_i \exp(\psi_0) - D_i^*]\exp(-\psi_0)$ in the test treatment group, where $D_i^* = \min(D_i, U_i \exp(\psi_0))$; and $T_i = D_i^* + [U_i - D_i^*]\exp(\psi_0)$ in the standard treatment group, where $D_i^* = \min(D_i, U_i)$. Finally, we set the observed failure time $X_i = \min(T_i, C_i)$, the observed censoring indicator $= I(X_i = C_i)$ and the observed time from randomization until treatment change $= \min(D_i^*, X_i)$, which is equal to the observed failure time for compliant cases.

To assess the sensitivity of the distributional assumption for ε_i, we repeated all of the above processes with a standard logistic distribution for ε_i in (2) instead of the extreme value distribution.

One weakness of g-estimation analysis is that it relies on the RPSFT model (1). In particular, the model incorporates a strong non-interaction assumption with respect to the treatment effect. To evaluate the sensitivity of the

treatment-by-covariate interaction in the g-estimation analysis, simulations were conducted under a true model of the two-parameter RPSFT model. The observed failure time T_i was calculated based on the following model:

$$U_i = \int_0^{T_i} \exp[-\psi_{10} A_i(t) - \psi_{20} A_i(t) I(L_{i1} \geqslant 0)]\, dt, \tag{3}$$

where $I(\cdot)$ is an indicator variable for the baseline covariate L_{i1}. The one-parameter model (1) was fitted using the extreme value distribution for ε_i in (2), in which (β_1, γ_1) were set to (1, 0) or (1, 1). For model (3), ψ_{10} was set to $-\log(\delta) = -\log(1.24)$ and ψ_{20} was set to -0.1 (hazard ratio among subjects with $L_{i1} \geqslant 0$ is 1.37) or -0.5 (hazard ratio among subjects with $L_{i1} \geqslant 0$ is 2.04).

We also conducted the simulations under the following interaction model:

$$U_i = \int_0^{T_i} \exp[-\psi_{10} A_i(t) - \psi_{20} A_i(t) I(L_{i1} < 0)]\, dt, \tag{4}$$

which means hazard ratio among subjects with $L_{i1} \geqslant 0$ is 1.24 and that of $L_{i1} < 0$ is 1.37 ($\psi_{20} = -0.1$) or 2.04 ($\psi_{20} = -0.5$). The true values for the overall hazard ratios in each simulation were calculated as the mean of the hazard had all subjects been continuously treated with the test treatment divided by the mean of the hazard had all subjects been continuously treated with standard treatment.

2.4. *Computation*

For the ITT and PP analyses, we used a Weibull regression analysis, because the RPSFT model (1) gives $T_i = U_i \exp(\psi_0)$, i.e. the accelerated failure time model $\log(T_i) = \psi_0 R_i + \log(U_i)$, if all patients comply with their allocated treatments throughout the study. For the PP analysis, two analysis data sets were used: a data set in which treatment-switching cases were censored at the time of the treatment change (PP1), and a data set in which treatment-switching cases were excluded from the analysis (PP2). Non-inferiority was assessed as described in Section 2.2 by calculating a $100(1-2\alpha)$ per cent two-sided confidence interval for the hazard ratio with $\alpha = 0.025$. For the g-estimation analysis, as described in Section 2.1, the g-test of $\psi_0 = -\log(\delta)$ (model (1)) or ψ_0 = true value for each simulation (model (3) or (4)) was conducted using the score test for the assigned treatment group R_i in the proportional hazards model. An SAS code to conduct the g-test is shown in Appendix. The proportion of significant cases at the one-sided α level was calculated as the empirical type I error. It is important to note that information on the baseline covariate L_{i1} was not used for any analysis.

3. Simulation results

The simulations were replicated 5000 times, giving a standard error on the estimated empirical type I error of 0.22 per cent. Table I shows the results of the simulations under the true structural model (1) and model (2) with an extreme value distribution for ε_i. The ITT analysis was anti-conservative for all simulations, except when independent treatment change was absent. The anti-conservativeness increased with the percentage of treatment change and was more pronounced for outcome-dependent treatment change. For outcome-dependent cases, anti-conservativeness was also seen when treatment change was absent, because the observed survival time depended on an unmeasured covariate, which can be considered as an omitted covariate in a Weibull regression model [27].

PP1 analysis (with treatment-switching cases censored at the time of the treatment change) maintained type I error near the nominal level of 0.025 for all levels of independent treatment change. For outcome-dependent cases, PP1 analysis was either conservative or anti-conservative depending on the mechanism underlying the percentage of treatment change. When treatment change in the test group was higher than in the standard group, PP1 analysis was anti-conservative, because more subjects in the test group with poor prognosis were censored at the time of treatment change; thus, the treatment effect was attenuated. There were cases for which PP1 analysis was more anti-conservative than ITT analysis at the same treatment change. On the other hand, when the treatment change in the test group was lower than in the standard group, PP1 analysis was noticeably conservative, because more subjects in the standard group with poor prognosis were censored at the time of treatment change; thus, a large bias in the treatment effect was observed. When the change in treatment was the same between treatment groups, type I errors were generally controlled below the nominal level.

PP2 analysis (with treatment-switching cases excluded from the analysis) was either conservative or anti-conservative even for independent treatment change. When the treatment change in the test group was higher than in the standard group, type I error was zero in all cases. On the other hand, when the treatment change in the test group was lower than in the standard group, PP2 analysis was noticeably anti-conservative. When treatment change was the same between treatment groups, type I errors were generally controlled below the nominal level. For outcome-dependent cases, similar

Table I. Empirical type I error (per cent) in model (2) with an extreme value distribution for ε_i (nominal error = 2.5 per cent)

Treatment change (per cent)		Independent treatment change				Outcome-dependent treatment change			
		Method				Method			
Standard	Test	ITT	PP1*	PP2†	G	ITT	PP1*	PP2†	G
0	0	2.70	2.70	2.70	2.56	4.72	4.72	4.72	2.44
	15	4.54	2.68	0.00	2.40	9.28	30.32	1.54	2.80
	30	7.30	2.60	0.00	2.46	15.46	70.08	0.48	2.36
	45	12.46	2.44	0.00	2.38	23.36	92.66	0.08	2.38
15	0	6.24	2.46	28.20	2.54	11.42	0.02	7.02	2.82
	15	8.78	2.50	2.00	2.54	18.72	2.60	3.02	2.34
	30	14.32	2.92	0.00	2.34	28.85	18.46	0.60	2.36
	45	20.38	2.72	0.00	2.42	38.46	53.90	0.10	2.48
30	0	10.72	2.44	84.40	2.68	19.38	0.00	10.46	2.56
	15	15.04	2.46	33.96	2.32	30.04	0.06	5.14	2.40
	30	24.22	2.48	2.02	2.50	42.18	1.64	1.84	2.66
	45	31.84	2.08	0.00	2.58	52.18	12.44	0.40	2.22
45	0	17.56	2.50	99.78	2.34	30.30	0.00	20.30	2.26
	15	26.06	2.50	92.80	2.64	42.02	0.00	11.06	2.38
	30	34.96	2.72	45.34	2.42	56.04	0.12	4.60	2.10
	45	45.98	2.42	1.98	2.54	66.24	1.78	0.94	2.32

*Treatment-switching cases were censored at the time of treatment change.
†Treatment-switching cases were excluded from the analysis.

results were obtained, although the degree of conservativeness or anti-conservativeness was smaller than for independent cases.

G-estimation analysis maintained type I error near the nominal level of 0.025 for all simulations. G-estimation analysis performed well even for outcome-dependent treatment change, although information on an unmeasured covariate was not used in the analysis.

Table II shows the results of the simulations under the true structural model (1) and model (2) with a standard logistic distribution for ε_i. PP1 analysis was anti-conservative for all simulations of independent cases. The anti-conservativeness was not dependent on the percentage of treatment change. On the other hand, g-estimation analysis maintained type I error near the nominal level of 0.025 both for independent and outcome-dependent cases. These results demonstrate the robustness of g-estimation analysis with the Weibull assumption for ε_i.

Tables III and IV show the results of the simulations under the true structural model of two-parameter RPSFT model (3) and (4), respectively. One-parameter g-estimation analysis was either anti-conservative (Table III) or conservative (Table IV) depending on the nature of the interaction. Because the simulation data were generated so that the larger the value of L_{i1}, the baseline failure time U_i is shorter, the result among subjects with $L_{i1} \geqslant 0$ will tend to affect the type I error. When the hazard ratios were larger among subjects with $L_{i1} \geqslant 0$ (model (3)), one parameter g-estimation analysis was anti-conservative for all simulations regardless of the percentage of treatment change. On the other hand, when the hazard ratios were smaller among subjects with $L_{i1} \geqslant 0$ (model (4)), conservativeness was observed throughout simulations. When the treatment-by-covariate interaction was small ($\psi_{20} = -0.1$), there were no large differences in the degree of anti-conservativeness or conservativeness between independent and outcome-dependent cases.

4. Discussion

In any randomized clinical trial, flaws in the design and conduct of the trial can lead to biased results. Bias resulting from violations of the entry criteria, non-adherence, missing data, or other deviations from the protocol tends to reduce sensitivity to treatment effects. This may be a more significant issue for non-inferiority trials, as it would tend to bias results toward a conclusion of similarity. In particular, treatment switching in non-inferiority trials is problematic. In this paper, we investigated the impacts of different types of treatment changes on the conservatism or anti-conservatism of analyses based on the ITT, PP, and g-estimation methods for a time-to-event outcome.

ITT results are anti-conservative in the presence of non-compliance due to treatment change as our simulations indicated. Thus, ITT results for non-inferiority trials must be carefully evaluated especially when the percentage of treatment change is high. However, the perception that ITT analysis biases results toward no difference may not always be true. When non-compliers have available a third treatment or no treatment, the effects of non-compliance may be

Table II. Empirical type I error (per cent) in model (2) with a standard logistic distribution for ε_i (nominal error = 2.5 per cent).

Treatment change (per cent)		Independent treatment change				Outcome-dependent treatment change			
		Method				Method			
Standard	Test	ITT	PP1*	PP2†	G	ITT	PP1*	PP2†	G
0	0	3.86	3.86	3.86	2.50	5.02	5.02	5.02	2.42
	15	5.68	3.44	0.02	2.26	9.98	21.28	0.70	2.52
	30	9.10	2.90	0.00	2.66	13.98	47.02	0.08	2.66
	45	14.26	2.90	0.00	2.72	19.72	72.08	0.00	2.22
15	0	6.96	4.18	35.54	2.66	10.72	0.40	15.96	2.28
	15	10.68	4.04	3.64	2.72	16.82	3.34	3.32	2.72
	30	13.94	3.62	0.02	2.58	23.58	14.74	0.48	2.78
	45	19.96	3.26	0.00	2.32	31.82	34.52	0.02	2.74
30	0	11.62	4.20	88.02	2.32	17.22	0.02	32.82	2.34
	15	16.16	3.88	40.00	2.56	24.50	0.34	10.68	2.46
	30	20.98	3.54	2.46	2.34	33.52	2.50	2.00	2.32
	45	28.92	3.82	0.00	2.58	42.34	9.80	0.40	2.50
45	0	17.56	4.36	99.82	2.64	23.00	0.00	58.08	2.38
	15	24.58	3.98	94.64	2.44	33.78	0.00	28.20	2.52
	30	29.64	3.68	49.78	2.16	43.00	0.34	8.74	2.22
	45	38.56	4.10	2.32	2.34	51.12	2.16	1.30	2.04

*Treatment-switching cases were censored at the time of treatment change.

†Treatment-switching cases were excluded from the analysis.

Table III. Empirical type I error (per cent) for one-parameter g-estimation analysis with true model (3) (nominal error = 2.5 per cent).

Treatment change (per cent)		Independent treatment change $(\beta_1, \gamma_1)=(1,0)$		Outcome-dependent treatment change $(\beta_1, \gamma_1)=(1,1)$	
Standard	Test	$\psi_{20}=-0.1$	$\psi_{20}=-0.5$	$\psi_{20}=-0.1$	$\psi_{20}=-0.5$
0	0	3.96	14.66	4.44	15.12
	15	3.40	13.26	3.12	15.16
	30	2.96	12.94	3.84	15.85
	45	3.66	11.32	4.50	14.68
15	0	3.84	12.86	4.38	15.02
	15	3.86	12.46	3.76	15.48
	30	3.34	11.20	3.84	15.48
	45	3.20	10.42	3.52	15.98
30	0	3.16	10.70	3.92	15.04
	15	3.58	11.00	3.94	15.38
	30	3.72	9.90	3.94	15.66
	45	2.88	8.36	4.28	15.52
45	0	3.56	9.78	4.10	14.38
	15	3.58	8.72	3.70	14.84
	30	3.18	8.78	3.80	15.00
	45	2.90	7.06	4.26	15.56

either in the conservative or anti-conservative directions [10, 15]. For example, ITT analysis may be conservative if the test and standard treatments are equally effective and clinicians tend to end the test treatment sooner than the standard one. In this case, the results for the test group will appear less effective in the ITT analysis when the non-compliers are not treated. Although ITT analysis is not ideal, it provides an estimate of the overall treatment effect that would be realized if the treatment was actually adopted and practiced in the community represented by the trial participants; thus, ITT analysis is important even for the non-inferiority trials.

PP analysis is often recommended for non-inferiority trials because of the perceived anti-conservativeness of ITT analysis. Our simulations indicated that PP analysis yields valid results only when treatment changes are independent of the outcome or when the percentages of treatment changes are the same between treatment groups. In the more likely scenario in which treatment changes depend on the outcome, PP analysis may be either conservative or anti-conservative depending on the mechanism underlying the percentage of treatment change. In general, we cannot know

Table IV. Empirical type I error (per cent) for one-parameter g-estimation analysis with true model (4) (nominal error = 2.5 per cent).

Treatment change (per cent)		Independent treatment change $(\beta_1, \gamma_1)=(1, 0)$		Outcome-dependent treatment change $(\beta_1, \gamma_1)=(1, 1)$	
Standard	Test	$\psi_{20}=-0.1$	$\psi_{20}=-0.5$	$\psi_{20}=-0.1$	$\psi_{20}=-0.5$
0	0	1.42	0.08	1.68	0.18
	15	1.86	0.18	1.50	0.10
	30	1.44	0.34	1.60	0.06
	45	2.16	0.20	1.48	0.12
15	0	1.40	0.28	1.44	0.12
	15	2.12	0.22	1.66	0.06
	30	1.72	0.30	1.34	0.08
	45	1.84	0.52	1.72	0.14
30	0	1.98	0.20	1.24	0.10
	15	1.70	0.34	1.36	0.08
	30	2.10	0.32	1.50	0.06
	45	2.04	0.40	1.48	0.12
45	0	1.74	0.38	1.96	0.20
	15	1.58	0.38	1.40	0.10
	30	1.94	0.34	1.34	0.06
	45	2.48	0.44	1.38	0.12

which mechanism is at work in typical applied settings. If treatment-switching cases are excluded from the analysis population (PP2 analysis), the approach could be seriously misleading, even for the case of independent change. Therefore, one needs to carefully define the PP population in the protocol. While PP analysis clearly addresses the actual benefits of intervention, it generally leads to biased results because the subjects who strictly adhere to the assigned treatment are usually a non-random sample of all study subjects. Therefore, even for non-inferiority trials, ITT analysis has an advantage over PP analysis, because the later fails to maintain the original randomized group.

Randomization-based g-estimation analyses are rarely used in practice. Our simulations showed that g-estimation analysis yields valid results even for outcome-dependent treatment change when the underlying structural model is correct. This approach maintains the original randomized group, whereas PP analysis does not. It also addresses the causal question of the actual benefits of intervention, which is a question of interest in the presence of non-compliance data, i.e. the effect that would be realized if all subjects complied with the treatment to which they were assigned. Therefore, g-estimation analysis should be more widely adopted, especially for non-inferiority trials. Greenland *et al.* [24] and Cole and Chu [28] stated that g-estimation should become a standard procedure for analysis of trials with non-compliance.

One weakness of g-estimation analysis is that it relies on the RPSFT model (1). As our simulations indicated, the approach was either conservative or anti-conservative when treatment-by-covariate interaction existed. The bias was generally in proportion to the degree of interaction. In practice, interactions with baseline covariates can be handled by stratification, i.e. subgroup analyses. These subgroup analyses will be required even in the ITT analysis when there exist important treatment-by-covariate interactions. More formally, Robins and Tsiatis [17] and Mark and Robins [18] suggested extension of the one-parameter RPSFT model to two-parameter models such as model (3). Multi-parameter RPSFT models have been fitted by Robins and Greenland [29] and White and Goetghebeur [30]. More work will be necessary to use the multi-parameter RPSFT models, especially when general non-compliance patterns that include treatment change, such as having no treatment or a third treatment, are observed in the non-inferiority trials. For example, when there are subjects who discontinue all treatments, two parameters have to be included in the RPSFT model; i.e. one is the effect of test treatment compared with no-treatment and the other is the effect of standard treatment compared with no-treatment. In this situation, there is a problem of one estimating equation for two unknown parameters. Further research will be needed for this issue.

Certain assumptions must be made to estimate the actual benefits of intervention in the presence of non-random non-compliance. Robins [16] suggests applying various analytical methods and discussing the consistency of the results and the assumptions underlying each method. Greenland *et al.* [24] also state that no single analysis is ideal and multiple analyses of the same data can be of benefit in interpreting study results, provided that the limitations of each analysis are recognized.

For simplicity, we assumed in the simulations that treatment change in each subject occurred only once during the study period. The RPSFT model (1), however, has no such actual limitation [17, 18, 23]. We also assumed that the time from randomization to the fixed end of the study without death was the only source of censoring. This assumption is

not realistic, because there will be drop-outs as well as administrative censoring. To adjust for selection bias due to non-administrative censoring, the inverse probability censoring weighted (IPCW) method has been proposed [31–34]. The underlying principle of the IPCW method is that estimation is based on the observed outcomes, but weights them to account for the probability of being uncensored. Finally, randomization-based g-estimation analysis can be performed for other types of outcomes, such as continuous, binary, or count responses, using structural nested mean models [20, 35].

In conclusion, the current requirement that non-inferiority be demonstrated for both the ITT and PP populations does not necessarily guarantee the validity of a non-inferiority conclusion. PP analysis should be phased out, and randomization-based g-estimation analysis should be used alongside ITT analysis, especially in non-inferiority trials.

Appendix A: SAS code for g-test and g-estimation

A g-test of the hypothesis $\psi_0 = -\log(\delta)$ can be conducted with standard survival analysis software by creating a new survival time (time_new) and event indicator (event_new) for each subject. These new variables are defined from the observed data and ψ_0 as follows:

```
/* t is survival time
c is administrative censoring time (time from randomization until the end of the study)
s_time is time from randomization until treatment switching (this variable is equal to the observed survival time in
a compliant case)
event = 1 if death, 0 if administratively censored
r = 1 if allocated to test treatment, 0 if allocated to standard treatment
psi = -log(delta) */
data g; set _data_;
if r=0 then ui=s_time+exp(-psi)*(t-s_time);
else if r=1 then ui=exp(-psi)*s_time+(t-s_time);
if event=1 then time_new=min(ui, c*exp(-psi), c); else time_new=min(c*exp(-psi), c);
if event=1 and time_new=ui then event_new=1;
else if event=1 and time_new ne ui then event_new=0;
else if event=0 then even_new=0;
```

/* A g-test of the hypothesis $\psi_0 = -\log(\delta)$ can be conducted using the score test for the assigned treatment group in the proportional hazards model as follows: */

```
proc phreg data = g; model time_new*event_new(0) = r; ods output GlobalTests=score;
run;
```

To estimate ψ_0 (the g-estimate), for each subject we calculate a value for the random variable $U_i(\psi)$ (the variable name 'ui' in the above SAS program) for each of a set $\{\psi\}$ of hypothesized values of ψ_0. For each ψ in the set $\{\psi\}$, we fit the above proportional hazards model with a new survival time (time_new) and event indicator (event_new), and tabulate the results. The particular ψ that yields a score test statistic of zero of r is the g-estimate of ψ_0. Although one should use a very fine search grid to find the estimate, the function of $\{\psi\}$ is not likely to be smooth; hence, we are unlikely to observe a single ψ with an associated test statistic of exactly zero. Therefore, we take as $\hat{\psi}$ the smallest ψ that changes the sign of the test statistic. A test-based 95 per cent confidence interval for $\hat{\psi}$ is obtained as the set of all ψ's with accompanying test statistic less than |1.96|.

Acknowledgements

I thank Masataka Taguri, PhD, for helpful comments. I am also grateful to the editor Vern Farewell, an associate editor, and two anonymous reviewers for their constructive comments on an earlier version of the paper. This work was supported by funding from the Health and Labour Sciences Research Grant for Clinical Cancer Research from the Ministry of Health, Labour and Welfare, Japan (the Study Group for Adjuvant Chemotherapy of Pancreatic Cancer).

References

1. Temple R, Ellenberg SS. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Annals of Internal Medicine* 2000; **133**:455–463.
2. Ellenberg SS, Temple R. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Annals of Internal Medicine* 2000; **133**:464–470.

3. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**:169–186. DOI: 10.1002/sim.1425.
4. Rothmann M, Li N, Chen G, Chi GTH, Temple R, Tsou HH. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine* 2003; **22**:239–264. DOI: 10.1002/sim.1400.
5. Hung HMJ, Wang SJ, O'Neill RT. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal* 2005; **47**:28–36. DOI: 10.1002/bimj.200410084.
6. Fleming TR. Current issues in non-inferiority trials. *Statistics in Medicine* 2008; **27**:317–332. DOI: 10.1002/sim.2855.
7. Ebbutt AF, Frith L. Practical issues in equivalence trials. *Statistics in Medicine* 1998; **17**:1691–1701.
8. Jones B, Jarvis P, Lewis J, Ebbutt A. Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* 1996; **313**:36–39.
9. ICH Harmonised Tripartite Guideline. E9: statistical principles for clinical trials, 1998.
10. Sanchez MH, Chen X. Choosing the analysis population in non-inferiority studies: per protocol or intent-to-treat. *Statistics in Medicine* 2006; **25**:1169–1181. DOI: 10.1002/sim.2244.
11. Committee on Proprietary Medical Products Points-to-Consider. Points to consider on switching between superiority and non-inferiority. CPMP, 2000.
12. Rohmel J. Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine* 1998; **17**:1703–1714.
13. Hauck WW, Anderson S. Some issues in the design and analysis of equivalence trials. *Drug Information Journal* 1999; **33**:109–118.
14. Garrett AD. Therapeutic equivalence: fallacies and falsification. *Statistics in Medicine* 2003; **22**:741–762. DOI: 10.1002/sim.1360.
15. Sheng D, Kim MY. The effects of non-compliance on intent-to-treat analysis of equivalence trials. *Statistics in Medicine* 2006; **25**:1183–1199. DOI: 10.1002/sim.2230.
16. Robins JM. Correcting for non-compliance in equivalence trials. *Statistics in Medicine* 1998; **17**:269–302.
17. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics* 1991; **20**:2609–2631.
18. Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial. *Controlled Clinical Trials* 1993; **14**:79–97.
19. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 1978; **6**:34–58.
20. Robins JM. Causal inference from complex longitudinal data. In *Latent Modelling with Applications to Causality*, Berkane M (ed.). Springer: New York, 1997; 69–117.
21. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009; **20**:880–883. DOI: 10.1097/EDE.0b013e3181bd5638.
22. Cox DR, Oakes D. *Analysis of Survival Data*. Chapman and Hall: London, 1984.
23. Hernan MA, Cole SR, Margolick J, Cohen M, Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety* 2005; **14**:477–491. DOI: 10.1002/pds.1064.
24. Greenland S, Lanes S, Jara M. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and g-estimation. *Clinical Trials* 2008; **5**:5–13.
25. Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika* 1995; **82**:835–845.
26. Korhonen PAK, Laird NM, Palmgren J. Correcting for non-compliance in randomized trials: an application to the ATBC study. *Statistics in Medicine* 1999; **18**:2879–2897.
27. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; **71**:431–444.
28. Cole SR, Chu H. Effect of acyclovir on herpetic ocular recurrence using a structural nested model. *Contemporary Clinical Trials* 2005; **26**:300–310. DOI: 10.1016/j.cct.2005.01.009.
29. Robins JM, Greenland S. Adjusting for differential rates of prophylaxis therapy for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association* 1994; **89**:737–749.
30. White IR, Goetghebeur EJT. Clinical trials comparing two treatment policies: which aspects of the treatment policies make a difference? *Statistics in Medicine* 1998; **17**:319–339.
31. Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate makers. *American Statistical Association Proceedings of the Biopharmaceutical Section*, Alexandria, VA, 1993; 24–33.
32. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
33. Robins JM, Finkelstein DH. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**:779–788.
34. Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine* 2009; **28**:1725–1738. DOI: 10.1002/sim.3585.
35. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics* 1994; **23**:2379–2412.

Overall Survival After Transarterial Lipiodol Infusion Chemotherapy With or Without Embolization for Unresectable Hepatocellular Carcinoma: Propensity Score Analysis

Kenichi Takayasu[1]
Shigeki Arii[2]
Iwao Ikai[3]
Masatoshi Kudo[4]
Yutaka Matsuyama[5]
Masamichi Kojiro[6]
Masatoshi Makuuchi[7]
for the Liver Cancer Study Group of Japan

Keywords: chemotherapy, hepatocellular carcinoma, iodized oil infusion, propensity analysis, transarterial chemoembolization

DOI:10.2214/AJR.09.3308

Received July 8, 2009; accepted after revision August 24, 2009.

[1]Department of Diagnostic Radiology, National Cancer Center Hospital, 5-1-1, Tsukiji, Chuo-ku, Tokyo 104-0045, Japan. Address correspondence to K. Takayasu.

[2]Department of Hepato-Biliary-Pancreatic Surgery, Tokyo Medical and Dental University, Graduate School of Medicine, Tokyo, Japan.

[3]Department of Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan.

[4]Department of Gastroenterology and Hepatology, Kinki University School of Medicine, Osaka, Japan.

[5]Department of Biostatistics, School of Health Sciences and Nursing University of Tokyo, Tokyo, Japan.

[6]Department of Pathology, Kurume University School of Medicine, Kurume, Japan.

[7]Department of Surgery, Japanese Red Cross Medical Center, Tokyo, Japan.

AJR 2010; 194:830–837

0361–803X/10/1943–830

OBJECTIVE. Although iodized oil transarterial chemoembolization (TACE) has been found to have survival benefit in the care of patients with unresectable hepatocellular carcinoma, iodized oil infusion chemotherapy without embolization has not been clearly found inferior to or equal to TACE. The purpose of this study was to determine whether one of these therapies is superior to the other or the two are equal in survival benefit and whether embolization with gelatin sponge particles is indispensable to prolonging survival.

SUBJECTS AND METHODS. A prospective nonrandomized observational cohort study was conducted over 8 years. Among 11,030 patients with unresectable hepatocellular carcinoma, 8,507 underwent TACE, and 2,523 underwent transarterial infusion therapy with an emulsion of iodized oil and an anticancer agent as initial treatment. Patients with extrahepatic metastasis or any previous treatment were excluded. The primary end point was all-cause mortality. To minimize selection bias, propensity score analysis was used to compare the two groups.

RESULTS. During the follow-up period, 5,044 patients (46%) died. In the analysis of all patients, TACE was associated with a significantly higher survival rate than infusion therapy without embolization (hazard ratio, 0.60; 95% CI, 0.56–0.64; $p = 0.0001$). The propensity score analysis showed that the hazard ratio for death in the TACE group ($n = 1,699$ patients) compared with the group who underwent infusion therapy without embolization ($n = 1,699$) was 0.70 (95% CI, 0.63–0.76; $p = 0.0001$). The median survival time of the TACE group was 2.74 years, and the 1-, 3-, and 5-year survival rates were 81%, 46%, and 25%. The corresponding values for the group who underwent transarterial infusion therapy without embolization were 1.98 years and 71%, 33%, and 16%.

CONCLUSION. Propensity score analysis showed that in the treatment of patients with unresectable hepatocellular carcinoma, TACE was associated with significantly better overall survival rates than was transarterial infusion therapy without embolization. TACE can be recommended as initial treatment of these patients.

Hepatocellular carcinoma (HCC) is the fifth most common type of cancer and the third most common cause of cancer mortality in the world [1]. The incidence of HCC is increasing in Japan [2], the United States [3], and other Western countries [4]. However, the number of patients who can undergo curative therapy such as resection, transplantation, and percutaneous ablation remains low. A 2005 report by the Liver Cancer Study Group of Japan showed transarterial chemotherapy, including transarterial chemoembolization with iodized oil and gelatin sponge particles (TACE) and transarterial iodized oil infusion chemotherapy without embolization, accounted for the initial treatment of 36.4% of 16,941 patients with HCC [5].

Randomized controlled trials [6, 7] and meta-analyses [8, 9] have shown that TACE is widely performed and recognized as having survival benefit in the treatment of patients with unresectable HCC accompanied by well-compensated cirrhosis. However, TACE is not always indicated, especially for patients with poor liver function and those with cancer in an advanced stage, because of the risk of hepatic failure and death after treatment [10, 11]. Instead, transarterial infusion therapy with an emulsion of iodized oil and an anticancer agent, also known as lipiodolization [12], has been performed for patients in poor condition [13–19].

A few reports have appeared on comparisons of the survival associated with transarterial iodized oil infusion therapy without

embolization and that associated with TACE, but no consensus has been reached. Two studies [18, 19] showed no significant difference between the two therapies, another study [14] showed infusion without embolization was associated with better survival than was TACE in a subgroup of patients at high risk, and another study [16] showed the reverse. We conducted a prospective nonrandomized observational cohort study to determine whether one of the therapies is superior to the other or whether the therapies are equal in survival benefit. We also evaluated whether gelatin sponge particles are indispensable to prolonging survival.

Subjects and Methods

Patient Characteristics

During the 8 years January 1994–December 2001, the Liver Cancer Study Group of Japan prospectively collected and biannually registered clinicopathologic data on 72,836 patients with primary liver cancer at nearly 800 medical institutions. Data were collected with a registration and questionnaire sheet with more than 180 questions. From that population, 11,030 patients (15.1%) with unresectable HCC were assigned to the current study cohort. Among these patients, 8,507 (77%) underwent TACE and 2,523 (23%) underwent iodized oil transarterial infusion therapy without embolization as initial treatment. These patients did not receive any other therapy during the first investigation period of no more than 2 years. Exclusion criteria were extrahepatic metastasis to lymph nodes and other organs and any previous treatment before the one studied. The 8,507 patients who underwent TACE in the current study were among 8,510 patients who participated in another study [20].

The diagnosis of HCC was based mainly on findings with imaging techniques such as sonography, dynamic CT, MRI, and angiography or on findings at pathologic study of biopsy specimens (4.7%). Abnormal elevation of levels of tumor markers also was found: α-fetoprotein greater than 400 ng/mL (normal, < 20 ng/mL) and des-γ-carboxyl prothrombin more than 100 mAU/mL (normal, < 40 mAU/mL). Typical HCC was visualized as high attenuation or signal intensity in the arterial phase and low attenuation or signal intensity or washout in the delayed phase (≈ 3 minutes after the initiation of contrast injection) of dynamic CT [21, 22] and dynamic MRI and as a hypervascular lesion at hepatic arteriography. Extrahepatic metastatic lesions were routinely examined with sonography, CT, and chest radiography.

The baseline characteristics of the 11,030 patients who underwent TACE (n = 8,507) and transarterial infusion therapy without embolization (n = 2,523) are shown in Table 1. The hepatic functional reserve was evaluated as liver damage in grade A, B, or C in the classification proposed by the Liver Cancer Study Group of Japan in 2000 and published in English in 2003 [23] (Table 2). This classification consists of five clinical and laboratory findings: ascites, serum bilirubin concentration, serum albumin concentration, indocyanine green retention rate at 15 minutes, and prothrombin activity. The severity of each clinical finding is evaluated separately. Degree of liver damage is based on the highest grade that contains at least two findings. This classification is closely related to the Child-Pugh classification and is more precise for discriminating whether patients with Child-Pugh A disease, that is, good candidates for surgical resection, have liver damage grade A or B [5, 24]. Concerning hepatitis B and C virus infection, four groups were categorized: negative result for hepatitis B virus surface antigen and positive result for hepatitis C virus antibody, positive result for hepatitis B virus surface antigen and negative result for hepatitis C virus antibody, positive results for both, and negative results for both. Maximum tumor size had four subgroups, and number of tumors had three subgroups.

Tumor Characteristics

The degree of vascular invasion of the portal vein consisted of the following four categories: Vp0, no invasion; Vp1, invasion to a third-order branch; Vp2, invasion to a second-order or segmental portal vein; and greater than Vp3, first-order portal vein including Vp4, main portal trunk. The degree of hepatic vein invasion was Vv0, no invasion, and greater than Vv1, any hepatic vein invasion, including the main hepatic veins and the inferior vena cava.

The TNM staging adopted in this study was proposed and revised by the Liver Cancer Study Group of Japan in 2000 (Table 3) and published in English in 2003 [23]. This revised TNM system was proposed as a new concordant TNM classification of primary liver cancer by the International Hepato-Pancreato-Biliary Association [25]. Namely, the T category is determined on the basis of the following three criteria: single lesion, tumor diameter 2 cm or less, and no vascular or biliary invasion (Table 3). Category T1 is determined when three criteria are fulfilled; T2, two criteria; T3, one criterion; and T4, no criteria. Stages I–IVA are determined mainly by the corresponding T category from T1 to T4.

Technique

A 5-French catheter was advanced to the superior mesenteric artery to confirm the patency of the portal vein trunk at postmesenteric portography. Common hepatic or celiac arteriography was performed to discern the number and location of lesions, tumor size, feeding artery, and presence of anatomic variation. A coaxial microcatheter (2.7 or 3.0 French) was selectively inserted through a 5-French catheter into the feeding artery as close to the lesion as possible. For multiple foci occupying the hepatic lobes, the right or left or both hepatic arteries were treated. For transarterial infusion therapy without embolization, an emulsion of iodized oil and an anticancer agent dissolved in contrast medium was injected with a three-way stopcock. For TACE, the emulsion was followed by injection of 0.5- to 1-mm-diameter gelatin sponge particles until cessation of blood flow was recognized under radiographic monitoring.

The following anticancer agents, in order of frequency used, were administered mostly as single agents but in some instances as part of multiple-drug therapy: doxorubicin (20–40 mg/m^2), epirubicin (30–60 mg/m^2), analogue of doxorubicin, mitomycin C, cisplatin, or zinostatin stimalamer (4–6 mg/kg body weight) [26]. The common dose of iodized oil was 5 mL/kg body weight (range, 3–10 mL). The entire dose of iodized oil and gelatin sponge particles was based on tumor size and the extent of the tumor. Follow-up consisted of dynamic CT or MRI with measurement of a tumor marker such as α-fetoprotein or des-γ-carboxyl prothrombin every 3–4 months. Therapy was repeated on demand when local recurrence (regrowth of the treated tumor), intrahepatic metastasis, or a second primary HCC was found and the patient would tolerate the therapy.

Statistical Analysis

The survival rates of patients who underwent TACE or transarterial infusion therapy without embolization were calculated from the date of diagnosis of HCC. Follow-up was ended on December 31, 2003. The primary end point was all-cause mortality. For the analysis of the patient characteristics of the TACE and therapy without embolization groups, chi-square or Mantel Trend chi-square tests were used. All-cause mortality was analyzed with univariate and multivariate Cox proportional hazards regression models.

Because this study was nonrandomized and observational, potential confounding (selection) bias was accounted for with propensity score analysis [27–29] and a multivariate Cox proportional hazards model. The propensity score is the probability that a patient with specific prognostic factors will receive treatment. It is a scalar summary of all observed prognostic factors. Within propensity score strata, prognostic factors in treated and control groups are similarly distributed, so that stratifying on propensity score strata removes overt selection bias due to the prognostic factors. We computed the propensity

TABLE 1: Baseline Characteristics of Patients With Unresectable Hepatocellular Carcinoma Who Underwent Transarterial Chemoembolization With Iodized Oil and Transarterial Iodized Oil Infusion Chemotherapy Without Embolization (n = 11,030)

Background Factor	Transarterial Chemoembolization With Iodized Oil (n = 8,507)		Transarterial Iodized Oil Infusion Chemotherapy Without Embolization (n = 2,523)		p
	No. of Patients	%	No. of Patients	%	
Age (y)					0.0144
< 60	1,845	22	604	24	
≥ 60	6,645	78	1,908	76	
Sex					0.4076
Men	6,120	72	1,836	73	
Women	2,385	28	686	27	
Degree of liver damage					< 0.0001
A	4,000	51	1,046	45	
B	3,052	39	964	41	
C	768	10	332	14	
Hepatitis B and C virus status					0.664
Hepatitis B surface antigen negative, hepatitis C virus antibody positive	6,063	74	1,795	74	
Hepatitis B surface antigen positive, hepatitis C virus antibody negative	895	11	266	11	
Both positive	212	3	58	2	
Both negative	972	12	311	13	
Maximum tumor size (cm)					0.0004
< 2	1,986	24	597	24	
2.1–3	1,980	24	577	24	
3.1–5	2,319	28	584	24	
> 5.1	2,072	25	684	28	
No. of tumors					0.0016
1	3,645	43	1,040	42	
2–3	2,676	32	689	28	
≥ 4	2,065	25	722	29	
Degree of portal vein invasion					< 0.0001
Vp0	6,881	88	1,777	77	
Vp1	322	4	90	4	
Vp2	305	4	130	6	
≥ Vp3	347	4	297	13	
Degree of hepatic vein invasion					< 0.0001
Vv0	7,246	97	1,936	95	
≥ Vv1	243	3	106	5	
α-Fetoprotein level (ng/mL)					< 0.0001
< 20	2,745	34	724	30	
21–400	3,393	42	994	41	
> 401	2,001	25	700	29	
TNM stage					< 0.0001
I (T1N0M0)	915	12	280	13	
II (T2N0M0)	2,908	39	719	34	
III (T3N0M0)	2,972	40	775	37	
IVA (T4N0M0)	639	9	318	15	

Note—Numbers in the sections do not equal those in the number columns because of missing values on the questionaire. Some percentages do not total 100 due to rounding.

TABLE 2: Degree of Liver Damage According to the Classification of the Liver Cancer Study Group of Japan

Clinical or Laboratory Finding	Grade of Liver Damage		
	A	B	C
Ascites	None	Controllable	Uncontrollable
Serum bilirubin concentration (mg/dL)	< 2.0	2.0–3.0	> 3.0
Serum albumin concentration (g/dL)	> 3.5	3.0–3.5	< 3.0
Indocyanine green retention rate at 15 minutes (%)	< 15	15–40	> 40
Prothrombin activity (%)	> 80	50–80	< 50

Note—Degree of liver damage is based on the highest grade containing at least two findings. For example, grade C applies if a patient has three clinical findings, one in column B and two in column C.

TABLE 3: Definitions of TNM Stage Proposed by the Liver Cancer Study Group of Japan

Classification	Criteria
T category	Single lesion, tumor diameter 2 cm or less, and no vascular or biliary invasion
T1	Fulfilling 3 criteria
T2	Fulfilling 2 criteria
T3	Fulfilling 1 criterion
T4	Fulfilling no criteria
TNM stage	
I	T1N0M0
II	T2N0M0
III	T3N0M0
IVA	T4N0M0, any T N1M0
IVB	Any T, N0–1M1

score by using multiple logistic regression with the dependent variable receiving TACE. The independent variables (prognostic factors) were the first nine variables (all but TNM stage) in Table 1.

To provide optimal control for confounding, propensity-based matching was used to select control patients similar to patients undergoing TACE. Using a macro (available at http://www2.sas.com/proceedings/sugi26/p214-26.pdf), we used propensity scores to match TACE patients to unique patients undergoing transarterial infusion therapy without embolization. We tried to match the background characteristics of the patient in the two groups by using propensity scores identical to five digits. If we could not make the match, we proceeded to four-, three-, two- and one-digit matches. We were able to match 1,699 TACE patients to 1,699 patients undergoing transarterial therapy without embolization.

For the 3,398-patient propensity score–matched sample, the survival curves were obtained with the Kaplan-Meier method and compared by log-rank test. Although performed with a nonrepresentative sample of patients undergoing treatment, matched analyses may yield a more valid estimate of treatment effect because patients with similar observed characteristics are compared, all of whom are candidates for selection of the treatment. All significance tests were two-tailed, and a value of $p < 0.05$ was considered statistically significant. All analyses were performed with statistical software (SAS version 9.1.3, SAS).

Results

Patient Characteristics in the Whole Sample

In the baseline characteristics of patients with unresectable HCC who underwent TACE ($n = 8,507$) and those who underwent iodized oil infusion chemotherapy without embolization ($n = 2,523$) (Table 1), there was a significant difference between the two groups in the following variables: age ($p = 0.0144$), liver function ($p < 0.0001$), maximum tumor size ($p = 0.0004$), number of tumors ($p = 0.0016$), portal and hepatic vein invasion ($p < 0.0001$), α-fetoprotein value ($p < 0.0001$), and TNM stage ($p < 0.0001$).

Crude Survival of TACE Patients and Patients Undergoing Therapy Without Embolization

During an 8-year follow-up period, 3,671 patients (43%) in the TACE group died, and data on the other 4,836 (57%) were censored; 1,373 patients (54%) in the therapy without embolization group died, and the data on 1,150 patients (46%) were censored. The median follow-up period was 1.39 years (range, 0.003–7.99 years) for the TACE group and 0.95 year (range, 0.003–7.97 years) for the therapy without embolization group. The median time and overall survival rates at 1-, 2-, 3-, 4-, 5-, and 7-years were 2.76 years and 82%, 62%, 46%, 34%, 25%, and 15% for the TACE group and 1.69 years and 66%, 45%, 31%, 23%, 15%, and 7% for the therapy without embolization group. There was a significant difference between two therapies (hazard ratio [HR], 0.60; 95% CI, 0.56–0.64; $p = 0.0001$).

Multivariate analysis of factors affecting time to death of patients who underwent TACE and iodized oil infusion chemotherapy without embolization showed that the following seven covariates were independent factors (Table 4): treatment (HR, 0.63; 95% CI, 0.59–0.68; $p = 0.0001$), degree of liver damage ($p = 0.0001$), maximum tumor size ($p = 0.0001$), number of tumors ($p = 0.0001$), portal vein invasion ($p = 0.0001$), hepatic vein invasion ($p = 0.001$), and α-fetoprotein value ($p = 0.0001$).

Survival of TACE Patients and Patients Undergoing Therapy Without Embolization Matched by Propensity Score

The baseline characteristics of 1,699 patients treated with TACE and 1,699 treated with transarterial iodized oil infusion chemotherapy without embolization matched by propensity score are shown in Table 5. Unlike the population as a whole, these two propensity-matched groups were well balanced. Regarding portal vein invasion, a significant difference seen among four subgroups was not seen in two subgroups categorized as Vp0–Vp1 and greater than Vp3.

The median follow-up periods for the TACE and infusion chemotherapy without embolization groups were 1.82 and 1.06 years, respectively. The patients with TACE had a lower risk of death than those who underwent treatment without embolization (HR, 0.70; 95% CI, 0.63–0.76; $p = 0.0001$). The median survival time and overall survival rates at 1-, 2-, 3-, 4, 5-, and 7-years were 2.74 years and 81%, 62%, 46%, 34%, 25%, and 15% for TACE versus 1.98 years and 71%, 49%, 33%, 23%, 16%, and 7% for therapy without embolization (Fig. 1).

Discussion

Infusion therapy of an emulsion of iodized oil and an anticancer agent without gelatin