Table 2
Influence of prognostic factors on overall survival

Prognostic factors		Number of patients	Influence of prognostic factors	
		n=2158	HR (p-value)[a]	HR (p-value)[b]
Sex	Male (%)	1304 (60.4%)	–	–
	Female (%)	854 (39.6%)	0.896 (p=0.1666)	0.866 (p=0.0713)
Age	≤49 (%)	376 (17.4%)	–	–
	50–59 (%)	687 (31.8%)	1.163 (p=0.1996)	1.241 (p=0.0673)
	60–69 (%)	783 (36.3%)	1.112 (p=0.3614)	1.176 (p=0.1626)
	≥70 (%)	312 (14.5%)	1.295 (p=0.0615)	1.418 (p=0.0117)
Histological depth of tumor invasion	≤pm (%)	556 (25.8%)	–	–
	ss/a1 (%)	758 (35.1%)	1.763 (p<0.0001)	1.434 (p=0.0037)
	≥s/a2 (%)	844 (39.1%)	2.873 (p<0.0001)	2.089 (p<0.0001)
Lymph node metastasis	n (−) (%)	1242 (57.6%)	–	–
	n1 (+) (%)	528 (24.5%)	2.159 (p<0.0001)	1.953 (p<0.0001)
	≥n2 (+) (%)	388 (18.0%)	3.632 (p<0.0001)	3.171 (p<0.0001)

[a] Crude hazard ratio and p-value using Cox regression.
[b] Adjusted hazard ratio by other three factors and p-value using Cox regression.

The results suggested that, as consistent with common knowledge in this field [12,13], histological depth of tumor invasion and lymph node metastasis had a particularly strong influence on overall survival.

3. Simulation methods

From the 2158 patients of the hypothetical population, 10,000 data sets of 50, 100, 150 and 200 patients (combining two groups) were repeatedly sampled with replacement to provide a data set for simulation. The sampling was conducted using the SURVEYSELECT procedure [14]. Allocation into two treatment groups, active (A) or placebo (P), was conducted repeatedly 10,000 times using three types of allocation methods: simple randomization, stratified randomization and minimization.

Simple randomization (SR) was conducted using pseudo Bernoulli random numbers. Two types of stratified randomization were performed: one is stratified randomization with four factors (STR4), sex, age, histological depth of tumor invasion and lymph node metastasis, and the other stratified randomization with the later two factors (STR2). In both cases, to ensure balance in the number of patients between groups within strata, the block size was set to four (an example of block AAPP). The influence of allocation probability and number of allocation factors on performance of minimization was evaluated, as well as balance in the simultaneous distribution of prognostic factors.

From the 10,000 sets derived for each trial size using the three types of allocation methods, 1000 sets were used to compare the differences in the number of patients between groups, balance in prognostic factors between groups and balance in the simultaneous distribution. The absolute value of the difference of the patient number between groups was the indicator for imbalance, and its 50 and 99 percentiles in the derived 1000 sets were evaluated, while the p-value of the chi-square test for the contingency table for each prognostic factor and the groups was calculated, and its 50 and 1 percentiles were used to compare the degree of balance among the three allocation methods. Interactions between prognostic

factors often arise in practical situations. Based on the features of allocation methods, stratified randomization is an allocation method that achieves balance in the simultaneous distribution of multiple prognostic factors, while minimization is an allocation method that achieves balance in the marginal distribution of each prognostic factor and does not ensure balance in the simultaneous distribution. Therefore, it has been suggested that if interactions exist among prognostic factors, stratified randomization is preferred over minimization [15]. Therefore, the p-value of the chi-square test of the contingency table formed by simultaneous distribution of multiple prognostic factors and group was calculated to evaluate balance in the simultaneous distribution.

Finally, the entire 10,000 sets of simulation data were used to evaluate the performance of statistical tests. Several statistical tests for actual overall survival time were conducted for evaluating the size of type I error after allocation, and the proportion achieving the chi-square test statistics greater than 3.841 (upper 5% point of chi-square distribution with df=1) was calculated to determine whether the nominal significance level (5%) was maintained. The applied tests were the log-rank test, the stratified log-rank test and the hazard ratio test using Cox regression. To compare the statistical power, based on an accelerated model in which the survival time increases or decreases due to the effect of treatment, overall survival is prolonged to 1.6- or 2.0-fold (censored at 5 years if it is longer than 5 years) for patients allocated to group A. In other words, if a patient died after 2 years from randomization, it was presumed that the patient died after 3.2 years ($2\times1.6=3.2$), and if a patient died after 4 years, it was presumed that the patient was censored at 5 years because it is longer than 5 years ($4\times1.6=6.4$). In special cases, that is, exponential and Weibull distributions, the accelerated model is equivalent to the proportional hazard model, and the hazard ratios are multiplied 1/1.6- or 1/2.0-fold, respectively.

4. Results

4.1. Imbalance in the number of patients between groups

The degree of imbalance in the number of patients between groups was compared among the three allocation methods. The allocation probability of minimization was changed from 1.00 (deterministic allocation) to 0.70 by 0.05. Table 3 shows the degree of imbalance in the number of patients between groups of 50-patient trials based on 1000 simulations.

Table 3
Degree of imbalance in the number of patients between groups of 50-patient trials based on 1000 simulations

Summary statistics	Absolute value of the difference in the number of patients between two groups									
	MIN (allocation probability)							STR2	STR4	SR
	1	0.95	0.90	0.85	0.80	0.75	0.70			
99 Percentile	2.0	2.0	2.0	4.0	4.0	4.0	4.0	10.0	16.0	18.0
50 percentile	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	4.0	4.0
Mean	0.5	0.6	0.7	0.8	0.9	1.0	1.1	2.9	4.9	5.6

Figures in the table show the summary statistics of the absolute value of the difference of patients between groups for each allocation method in 1000 simulations (N=50).
MIN: minimization included all four factors; STR2: stratified randomization included two factors; STR4: stratified randomization included all four factors; SR: simple randomization.

Using the SR, the 50 percentile was 4 (in actual case A=27 and P=23), and the 99 percentile (in the worst scenario) was 18 (A=34 and P=16).

The 50 percentile was 2, and the 99 percentile was 10 using the STR2. On the other hand, the 50 and 99 percentiles of STR4 were 4 and 16, respectively. In this case, the number of strata 72 (2×4×3×3) was relatively large, compared with trial size 50, therefore, a decrease in the number of patients per stratum resulted in large differences in the number of patients between groups.

Decreasing the allocation probability of minimization provided a slightly worse balance in the number of patients; however, even with the allocation probability of 0.70, balance still remained good, contrasting to the STR2, the STR4 and the SR. The difference in the number of patients between groups using minimization with the allocation probability of 0.70 was 0 (50 percentile) and 4 (99 percentile).

Although similar tendencies were confirmed with trial sizes of 100, 150 and 200 patients, the degree of imbalance of the SR, the STR2 and the STR4 was improved. The 50 and the 99 percentiles are shown in Table 4.

4.2. Balance in prognostic factors

Table 5 shows the degree of imbalance in prognostic factors in the case of 50-patient trials based on 1000 simulations.

Using the SR, under the null hypothesis, the p-value is expected to be uniformly distributed between 0 and 1; therefore, results of the simulation showed that the 50 percentile of p-values for all prognostic factors was around 0.50, and similarly, the 1 percentile of p-values was about 0.01.

It was confirmed that smaller allocation probability in minimization offered greater imbalance in prognostic factors.

Balance in the allocation factors achieved with the STR2 was similar to that achieved with minimization at the allocation probability of 0.70; however, balance in nonstratified factors such as sex and age were comparable with that obtained from the SR. In contrast, the STR4 had nonignorable imbalance, due to too many strata compared with a given trial size. It is difficult to apply stratified randomization to achieve good balance in such small-scale trials.

When selecting the allocation probability in minimization, if the 1 percentile of p-value for the chi-square test requires about 0.50 as a criterion to achieve a strongly acceptable degree of

Table 4
Degree of imbalance in the number of patients between groups of 100- to 200-patient trials based on 1000 simulations

Trial size	Absolute value of the difference in the number of patients between two groups							
	MIN0.7		STR2		STR4		SR	
	50 Percentile	99 Percentile	50 Percentile	99 Percentile	50 Percentile	99 Percentile	50 Percentile	99 Percentile
100	2	4	2	10	6	20	6	26
150	2	4	2	10	6	23	8	30
200	0	4	2	10	6	24	10	35

Figures in table show 50 and 99 percentiles of the absolute value of the difference of patients between groups for each allocation method, in 1000 simulations (N=100, 150 and 200, respectively).
MIN0.7: minimization included all four factors (allocation probability=0.70); STR2: stratified randomization included two factors; STR4: stratified randomization included all four factors; SR: simple randomization.

Table 5
Balance in the marginal distribution of prognostic factors of 50-patient trials based on 1000 simulations

Prognostic factor	MIN (allocation probability)							STR2	STR4	SR
	1	0.95	0.90	0.85	0.80	0.75	0.70			
50 Percentile of p-value of chi-square test for the contingency table of each factor and group										
Sex	0.7815	0.7766	0.7773	0.7766	0.7766	0.7745	0.7708	0.4741	0.5557	0.5218
Age	0.9438	0.9381	0.9297	0.9225	0.8975	0.8732	0.8397	0.4919	0.5745	0.4919
Histological depth of tumor invasion	0.9344	0.9331	0.9279	0.9256	0.9098	0.8629	0.8387	0.8151	0.5909	0.4948
Lymph node metastasis	0.9195	0.9162	0.9082	0.8905	0.8559	0.8437	0.8226	0.7702	0.5331	0.4723
1 Percentile of p-value of chi-square test for the contingency table of each factor and group										
Sex	0.5443	0.5107	0.3954	0.3705	0.3737	0.2575	0.2410	0.0138	0.0225	0.0121
Age	0.5086	0.5019	0.4725	0.3728	0.2917	0.2647	0.2309	0.0140	0.0265	0.0139
Histological depth of tumor invasion	0.6125	0.5831	0.5242	0.4897	0.4415	0.3236	0.2971	0.2789	0.0267	0.0109
Lymph node metastasis	0.5318	0.4867	0.5023	0.3621	0.2903	0.2275	0.2525	0.1405	0.0275	0.0141

Figures in table show 50 and 1 percentiles of p-value of chi-square test from 1000 simulation data sets (N=50) for each allocation method.
MIN: minimization included all four factors; STR2: stratified randomization included two factors; STR4: stratified randomization included all four factors; SR: simple randomization.

balance, even in the possible worst case, the allocation probability should be set to 0.95 in 50-patient trials.

Similar balance ranking was observed among the allocation methods in trials with 100, 150 and 200 patients; however, the degree of imbalance of the STR2 and the STR4 improved as trial size increased. Based on the above criterion 1 percentile of p-value is about 0.50, the allocation probability required for a given trial size would be 0.80 for 100, 0.75 for 150, and 0.70 for 200 patients. If the number of patients increases, it is possible to use a smaller allocation probability to avoid predictability, while keeping a good balance.

4.3. Balance in simultaneous distribution

The difference of 3×3 simultaneous distribution (histological depth of tumor invasion and lymph node metastasis) between groups were examined using the p-value of the chi-square test as an indicator, and the 50 and the 1 percentiles of p-value were calculated. Table 6 shows the results of 200-patient trials based on 1000 simulations.

As expected, the STR2 achieved the best comparability in the combination (nine levels) of histological depth of tumor invasion and lymph node metastasis. Minimization did not provide a good balance in simultaneous distribution, even when increasing allocation probability or the trial size.

Table 7 shows a typical pattern in which minimization did not work well; that is, marginal distribution was comparable between groups. However, nonignorable imbalance was observed in the simultaneous distribution. On the other hand, stratified randomization guarantees a good balance in the simultaneous distribution.

Fig. 1 shows the hazard ratios for each level in 2158 patients of the hypothetical population.

Histological depth of tumor invasion correlated with lymph node metastasis, but there was no strong interaction in the overall survival (Wald test, chi-square=3.086, df=4). If interaction cannot be ignored, it is important to ensure balance in the simultaneous distribution. To achieve balance in the simultaneous distribution similar to stratified randomization, minimization should be applied by combining these two factors into one allocation factor with nine levels.

Table 6
Balance in simultaneous distribution of 200-patient trials based on 1000 simulations

Simultaneous distribution[a]	p-Value of chi-square test for the contingency table of simultaneous distribution and groups									
	MIN (allocation probability)							STR2	STR4	SR
	1	0.95	0.90	0.85	0.80	0.75	0.70			
50 Percentile	0.8882	0.8748	0.8869	0.8909	0.8873	0.8677	0.8642	0.9973	0.8665	0.4741
1 Percentile	0.1497	0.0837	0.1048	0.1192	0.1127	0.1123	0.1072	0.8569	0.2331	0.0081

Figures in table show 50 and 1 percentiles of p-value of chi-square test from 1000 simulation data sets (N=200) for each allocation method.

MIN: minimization included all four factors; STR2: stratified randomization included two factors; STR4: stratified randomization included all four factors; SR: simple randomization.

[a] Combinations (nine levels) of the depth of tumor invasion and the lymph node metastasis.

Table 7
Balance between groups in the combination of levels of a 200-patient trial

		Lymph node metastasis			Group A	Group P
		n (−)	n1 (+)	≥n2 (+)		
STR2						
Histological depth of	≤pm	44 [A: 23, P: 21]	2 [A: 1, P: 1]	10 [A: 6, P: 4]	29	27
tumor invasion	ss/a1	34 [A: 17, P: 17]	21 [A: 10, P: 11]	8 [A: 4, P: 4]	31	32
	≥s/a2	45 [A: 22, P: 23]	18 [A: 9, P: 9]	18 [A: 9, P: 9]	40	41
Group A		61	20	19	100	
Group P		62	21	17		100
Min0.70						
Histological depth of	≤pm	44 [A: 23, P: 21]	2 [A: 2, P: 0]	10 [A: 2, P: 8]	27	29
tumor invasion	ss/a1	34 [A: 12, P: 22]	21 [A: 13, P: 8]	8 [A: 7, P: 1]	32	31
	≥s/a2	45 [A: 26, P: 19]	18 [A: 6, P: 12]	18 [A: 8, P: 10]	40	41
Group A		61	21	17	99	
Group P		62	20	19		101

Figures in table show the example of result of allocation from one of 1000 simulations (N=200) for each allocation method for STR2 and MIN0.7. In brackets is the number of allocated patients in each group.
MIN0.7: minimization included all four factors (allocation probability=0.70); STR2: stratified randomization included two factors.

4.4. Type I error and power

Stratified log-rank test and Cox regression were conducted adjusting for two factors: histological depth of tumor invasion and lymph node metastasis. Table 8 shows the results of 200-patient trials based on 10,000 simulations.

When stratified randomization or minimization was used, it was apparent that the result of unadjusted test for allocation factors (log-rank test) turned out to be conservative. In contrast, analysis with adjustment for the allocation factors as covariates such as stratified log-rank test and Cox regression, provided type I error close to the nominal significance level and, as a result, improved the statistical power. Parallel affinities were noticed with trial sizes of 50, 100 and 150 patients.

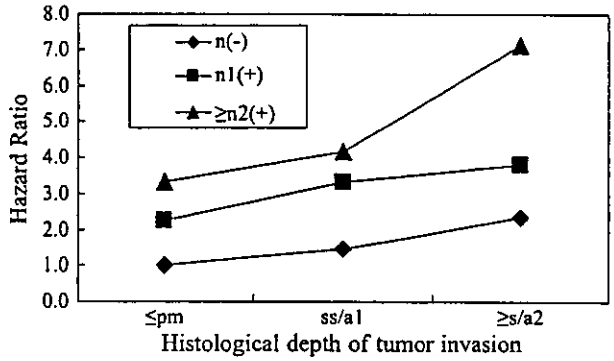


Fig. 1. Hazard ratio (n (−), pm or less as 1) for each level of combination (nine levels) of histological depth of tumor invasion and lymph node metastasis in 2158 patients. Histological depth of tumor: ≤pm: m, sm and pm; ≥s/a2: s/a2 and si/ai. Lymph node metastasis: ≥n2 (+): n2 (+), n3 (+), and n4 (+).

Table 8
Type I error and power of 200-patient trials based on 10,000 simulations

Active group (×survival time)		Test	Allocation methods		
			MIN (detaministic)	STR2	SR
H_0	(×1.0)	Log-rank	0.0352	0.0334	0.0505
		Stratified log-rank	0.0466	0.0446	0.0504
		Cox regression	0.0494	0.0475	0.0548
H_1	(×1.6)	Log-rank	0.4155	0.4265	0.4302
		Stratified log-rank	0.4790	0.4904	0.4778
		Cox regression	0.4799	0.4853	0.4812
	(×2.0)	Log-rank	0.7793	0.7852	0.7610
		Stratified log-rank	0.8215	0.8303	0.8054
		Cox regression	0.8220	0.8271	0.8118

The table shows the actual type I error at a nominal significance level of 0.05. The table also shows the power (H_1) attained with the overall survival prolonged for allocated to group A. Both cases in the 10,000 simulations (N=200).
MIN: minimization included two factors (allocation probability=1); STR2: stratified randomization included two factors; SR: simple randomization.

It is evident that these results could be predicted qualitatively based on the feature of the analysis methods; however, this study confirmed the extent of the difference among allocation methods in a quantitative manner, based on actual clinical trial data. The simulation also revealed that type I error of the minimization could be sufficiently maintained both with and without adjustment analysis. This suggested the statistical validity of minimization.

5. Discussion

We conducted simulations using the actual data from clinical trials of rectal cancer and they provided the following results:

(1) When four allocation factors exist, stratified randomization does not perform well in small-scale trials (about 50 patients). However, even in such cases, minimization can achieve sound balance in the number of patients and in the distribution of prognostic factors between groups. It can be concluded that the results indicate the usefulness of minimization, which can achieve balance even in smaller scale trials.
(2) Minimization can ensure comparability between groups even using smaller allocation probability (instead of deterministic allocation) to prevent predictability by increasing the sample size.
(3) Minimization can ensure balance in the marginal distribution of prognostic factors but does not ensure balance in the simultaneous distribution. Therefore, minimization can be difficult if interactions between prognostic factors can be predicted. In such cases, stratified randomization or minimization in which levels are reclassified based on simultaneous distribution of multiple prognostic factors should be applied.
(4) Simple unadjusted tests have conservative type I errors when minimization is conducted. On the other hand, because adjusted tests for allocation factors as covariates can achieve an approximate nominal significance level, they have the elevated power by up to 5–6%.

Three main criteria were used to assess performance of each allocation method:

(1) balance in the number of patients between treatment arms,
(2) balance in the distribution of prognostic factors,
(3) performance of statistical tests.

The reason why balance in the number of patients allocated to each treatment arm is desirable in a randomized trial is that, for a fixed number of patients, statistical power is maximized and the width of confidence interval is minimized when an equal number of patients is allocated to each arm, although moderate imbalances produce negligible loss of statistical performance.

As confirmed in our study, if known prognostic factors are balanced in the allocation process by means of stratification or minimization, it is obvious that these prognostic factors are more evenly distributed than with simple randomization, especially in small trials. Assuring a balance in the distribution of prognostic factors provides an unbiased estimation of hazard ratio and the rationale for the use of simple statistical methods without adjusting for prognostic factors.

When marked imbalances are found, these can be adjusted in the analysis; however, a variety of possible models are available, according to the difference of mathematical formulation and combination of prognostic factors. Moreover, the validity of adjusted analysis depends on the correctness of model assumptions that cannot be confirmed. Therefore, it is much better to balance major prognostic factors at the design stage and to apply an unadjusted simple statistical method, although, even in the balanced case, adjusted analysis gives greater power.

In this study, small-scale clinical trials were evaluated. The total number of patients was between 50 and 200 in the two groups. In actual comparative cancer clinical trials, the trial size is often larger than these sizes, especially in the adjuvant setting, cancer trial usually imply the enrollment of many hundreds (and sometimes more than thousand of patients) when endpoint is survival. However, if the trial size is large enough, simple randomization works sufficiently to obtain a good balance, and special allocation techniques are not required. And cancer trials assessing biological treatments potentially associated with dramatic effects on advanced disease or using the other endpoint, such as event-free survival, QOL and response rate, can require much fewer patients. In addition, it is usually required to conduct interim analysis in long-term and large-scale cancer clinical trials. Interim analysis is often conducted at the time when a third or a half of planned total event is accumulated. Therefore, sample size at the interim analysis is much smaller than that at the final analysis. Moreover, it is very important that the decision-making process is based on only minimum information to avoid any possible biases when the results of interim analysis are leaked. In addition, decision making based on interim analysis must be conducted in a very limited time frame.

Therefore, it is preferable to achieve strict balance in the number of patients and prognostic factors between groups at the time of the interim analysis to avoid complex analysis, such as adjusted analysis. Even in large-scale trials, it is important to ensure balance at the time of interim analysis.

Although four allocation factors were considered in our study, in some clinical trials, more allocation factors must be considered. When more allocation factors exist, other investigators have evaluated the performance of minimization in trial sizes of 1000 patients and prognostic factors involving 12 variables, and they demonstrated that minimization achieved balance and maintained nominal significance level

[16]. The study used actual stroke-patient data; however, the sampling method from the population was different from our study.

The covariate, age, was dealt as a categorical variable after categorizing into four levels in minimization of our study. However, age is essentially measured in a continuous way. Extensions of minimization to balance of the means and standard deviations of continuous prognostic factors between groups have also been proposed [17,18]. If a prognostic factor is a continuous variable and clinically appropriate categorization is difficult, the minimization which provides balance of the means and standard deviations are the method of choice.

Finally, the problem of dynamic allocation, such as minimization, as indicated by CPMP is discussed, based on our simulation results.

Predictability is one of the most important issues in the conduct of clinical trials. However, even when using reducing allocation probability, it is possible to achieve good balance between groups in a moderate-scale trial; therefore, this does not directly obstruct the application of minimization. It is indicated that balance in prognostic factors and the number of patients can be achieved in small clinical trials and that the significance level can be maintained, thus, the clinical and statistical justification of minimization was demonstrated. Indication by CPMP that "when dynamic allocation is used, the allocation factors should be considered in the analysis" was confirmed to be appropriate. The results of the simulation revealed that adjustments for the allocation factors can bring closer to the nominal significance level, with a pay-off being an improvement in the power by about 5%. Thus, when using minimization, it is necessary to specify the adjusted method in the statistical analysis plan.

In this investigation, where we used the data from rectal cancer patients in clinical trials of other fields, there are often multiple prognostic factors that reflect the severity of disease, corresponding to the histological depth of tumor invasion and lymph node metastasis investigated here, besides basic demographic variables such as sex and age. In conclusion, this investigation demonstrated that in small-scale clinical trials where multiple prognostic factors exist, minimization is a useful method to achieve balance in prognostic factors.

Acknowledgments

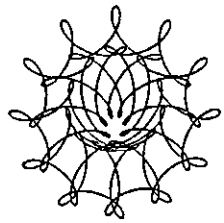
We are grateful to Wataru Kashiwagi (Taiho Pharmaceutical) and two anonymous referees for their useful comments. This research was (partially) supported by JSPS Grant-in-Aid for Scientific Research No. 16200022.

References

[1] Rosenberger WF, Lachin JM. Randomization in clinical trials, theory and practice. New York: John Wiley & Sons; 2002.
[2] Taves DR. Minimization: a new methods of assigning patients to treatment and control groups. Clin Pharmacol Ther 1974;15:443–53.
[3] Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 1975;31:103–15.
[4] ICH E9 Expert Working Group. Statistical principles for clinical trials. Stat Med 1999;18:1905–42.
[5] Japan Pharmaceutical Manufacturers Association. Drug allocation in clinical trials. Jpn Pharmacol Ther 1991;19: 2995–3013 [in Japanese].

[6] Pocock SJ. Clinical trials: a practical approach. New York: John Wiley & Sons; 1984.
[7] Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials: a review. Control Clin Trials 2002;23:662–74.
[8] McEntegart DJ. The pursuit of balance using stratified and dynamic randomization techniques: an overview. Drug Inf J 2003;37:293–308.
[9] The European Agency for the Evaluation of Medicinal Products. Points to consider on adjustment for baseline covariates. CPMP/EWP/2863/99 London, 2003. ⟨⟨http://www.emea.eu.int/htms/human/ewp/ewpptc.htm⟩⟩. Accessed December 11, 2003.
[10] Japanese Society for Cancer of the Colon and Rectum. General rules for clinical and pathological studies on cancer of the colon, rectum and anus. Sixth edition. Tokyo, Japan: Kanehara and Co.; 1998. In Japanese.
[11] Dukes CE. The classification of cancer of the rectum. J Pathol 1932;35:323–32.
[12] Minsky BD, Mies C, Recht A, Rich TA, Chaffery JT. Resectable adenocarcinoma of the rectosigmoid and rectum: 1. Patterns of failure and survival. Cancer 1988;61:1408–16.
[13] Minsky BD, Mies C, Recht A, Chaffery JT. Potentially curative surgery of the colon cancer: 1. Patterns of failure and survival. J Clin Oncol 1988;6:106–18.
[14] SAS Institute. SAS/STAT User's Guide, Version 8.
[15] Tu D, Shalay K, Pater J. Adjustment of treatment effect for covariates in clinical trials: statistical and regulatory issues. Drug Inf J 2000;34:511–23.
[16] Weir CJ, Lees KR. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. Stat Med 2003;22:705–26.
[17] Nishi T. Allocation to achieve balance in the means between treatment groups. Proceedings of the Biometric Society of Japan symposium; 2002. p. 43–8. In Japanese.
[18] Takaichi A, Nishi T. The allocation of patients using the extended minimization method. Proceedings of the 21th annual SAS Users Group International Japan conference; 2002. p. 467–71. In Japanese.

臨床統計学
臨床試験を中心として



大橋 靖雄

臨床統計学（Clinical Biostatistics）は，健康・医療に関わる応用統計学 Biostatistics（生物統計学と訳される）の重要な1分野である．わが国でも生物統計学の重要性がようやく認識され，実務家・研究者の需要が立ち上がるとともに教育システムも生まれつつある．大学では 1992 年に東京大学に初めての講座が誕生し，1999 年の北里大学薬学系研究科，2000 年の京都大学医学系研究科と講座開設が続いた．2002 年には東京理科大学大学院（経営工学）に社会人対象修士コースが生まれ，2004 年には久留米大学，厚生労働省保健科学院にも教育コースが開設予定である．現時点で生物統計家の参画が最も必要とされているのは，臨床研究とくに臨床試験の計画と解析である．本稿では，歴史的な展開も含め，詳細な数理というよりは，このような専門家が必要となってきた背景の記述と主要な概念について述べてみたい．

1. 臨床医学と EBM，臨床試験

臨床医学の目的は，患者の疾患を正確に診断し適切に治療を行うことにある．治療目的は，可能な場合には治癒を目指し，治癒が不可能な場合には患者の QOL を可能な限り向上させるか維持することである．しかし，医療提供者側の知識・技術の不完全さ，主に患者・疾患の多様さに多く由来する治療効果の不確実さから，診断結果や治療結果には永遠に除去不可能な曖昧さが伴う．したがって，診断・治療によってもたらされるベネフィットと被るリスクは，ともに「可能性」として確率変数的性格を帯びる．これらに対する患者の重み付けも患者個々の価値判断を反映して異なるはずのものであり，診断・治療法の選択は，リスク・ベネフィット両者のバランスと資源の制約の中で，本来は充分な情報提示と理解，そして自発的意志を前提としたインフォームドコンセントによるべきである（とされる）．しかし，医療提供者と患者の有する情報の不均衡，意志・患者双方の意識の問題もあり，これまでの治療上の意志決定は，医師主導のパターナリズムの中で，曖昧な状況下で行われてきたといってよい．近年の情報公開あるいは患者の権利主張の流れは，このような意思決定プロセスに大きな変革を与えようとしている．厚生労働省は，ここ数年，疾患毎に標準治療をまとめたガイドライン策定を各関連学会に依頼し，数多くの疾患ガイドラインやその案が発表されてきた．これは，国民皆保険のもと，治療結果をマスとして評価するシステムが存在しないまま，医師の自由裁量のもと出来高払いで治療が行われてきた反省と，また上記の情報開示の要求に応えるためでもあった．そして，このガイドライン策定過程で（医療関係者には周知であったが）改めて次の事態が浮き彫りになった．

「わが国には客観的証拠（evidence）が無い！」

臨床医学系学会では，最近5年ほど Evidence Based Medicine（略して EBM）という言葉が大流行である．EBM とは，目の前の患者の問題点を一定の手順で定型化し，主に文献検索と抽出された文献の批判的吟味により過去の「証拠・根拠」を点検し，そこから有効な情報を引き出し，目の前の患者に対して，その特異性を充分考慮しつつ，ときにはこれまでの経験によって修正も加え実践することである．科学と経験の融合である．ここで，evidence を提供するのが患者を対象とした臨床研究成果である．

臨床研究は，患者の診断・治療経過をまとめた症例報告，多数の患者の治療実態下での観察研究，そして臨床試験に大まかに分けられる．薬・手術・放射線などの治

療（予防）手段を研究者側が前向きに制御して行われる実験的研究が臨床試験である．これらの手段をまとめて介入（intervention）と呼ぶ．複数の臨床試験を統計的に併合する「メタアナリシス（meta-analysis）」も，もとの臨床試験の質が高ければ，高い質の evidence を提供し，ガイドライン策定に大いに利用されている．

臨床試験はヒトを対象とした実験ではあるが，ナチスあるいはわが国の石井部隊（731 部隊）が戦時中に行ってきた人体実験とは，倫理性の確保という点で大いに異なる．このための国際合意の実施基準が後述の GCP（Good Clinical Practice）である．

さて，治療法の開発とくに新薬開発に関して一般国民が抱いている最大の誤解は，画期的な新薬が動物実験や試験管内の基礎研究から生れる，という「幻想」である．すべての新薬の認可あるいは既存薬剤の適応拡大には，適切に計画・管理実施され，通常は適切な比較対照を有する（これを adequately well-controlled と呼ぶ）臨床試験が必須である（この事情は医療機器についても同様であるが，簡単のため以下は「薬」「製薬会社」で代表させる）．

GCP に従った臨床試験は，図 1 に示すように入念な準備と複雑な手順，多くの人々の参加・協力により実施される．事前にプロトコルと呼ばれる研究計画書が策定される．これは研究の意義（科学的側面と倫理的側面）と実施マニュアルの両面を持つ基本文書であり，この中で試験の目的・対象者・介入手段・統計解析を含む評価方法が規定される．このプロトコルに従って試験が進行するために，研究スポンサー（後述する「治験」では製薬会社）あるいは実施施設（病院）は手順書（Standard Operational Procedure; SOP）を定め，人的配置を行う．製薬会社や臨床試験の請負会社（Contract Research Organization; CRO）の社員が施設を訪問して実施状況と報告の正確さを点検する活動がモニタリングであり，さらに第三者の監査等を通じて，得られたデータの信頼性の品質保証がなされる．統計解析を担当する統計家は計画段階から参画し，解析計画書を策定し解析実務を行い，膨大な量となる報告書作成に協力する．

新薬申請の場合は，このような臨床試験を通じて当該薬剤が安全性の点で許容可能であり，さらにもちろん患者の生存・QOL やこれらと直結していることが証明されている評価変数（エンドポイントと呼ばれる）の統計的処理を通じて有効であることの証明がなされて初めてデータが審査当局（日本の場合は厚生労働省）に提出される．そして，1 年ほどの専門家による審査を経て，認可・発売となる．新薬 1 薬剤の開発費用はうなぎ上りであり，失敗例も含めれば 1 薬剤 100–200 億円以上，うち臨床試験の費用は数 10 億円以上となる．数千例を超えるような大規模試験の場合には数 100 億円の費用を要することもある．

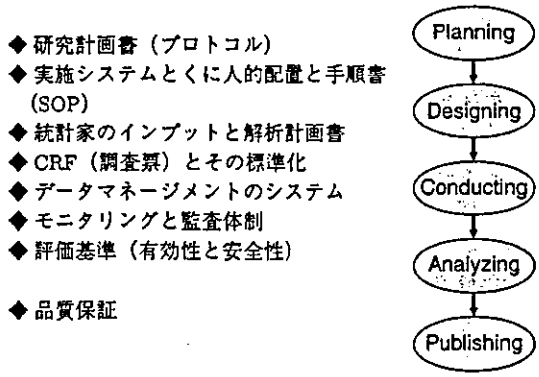


図 1　臨床試験に必要な準備．

2. わが国の evidence 不足

さて EBM に戻ろう．わが国の臨床医学は，evidence 作り，とくに臨床試験を通じた evidence に実はほとんど寄与していない．その理由は，以下のような，臨床試験実施のためのインフラストラクチャを欠いていたことにある．

(1) 医師研究者を支援するコーディネータ（Clinical Research Coordinator）
(2) 研究計画書（プロトコル）を書くことのできる医師研究者とそのための教育
(3) 生物統計家（試験統計家）あるいは生物統計学の教育
(4) 効率的なデータマネージメント・システムと中立的なデータセンター
(5) Peer review あるいは監査・査察によるデータの品質保証体制
(6) 医師研究者主導型の臨床研究に対する法的規制
(7) 結果を発信するプロフェッショナルである Medical writer の人材と教育
(8) 臨床試験の意義の理解と患者参加に対するインセンティブ

(9) 臨床試験に参加する医師研究者へのインセンティブ

(10) 研究資金とくに公的研究資金の供給と評価

「治験」とは，製薬会社がスポンサーとなり新薬申請あるいは適応拡大申請のために行われる臨床試験である．臨床試験には，製薬会社のスポンサーで市販後に行われる（通常は大規模な）市販後臨床試験や，公的研究費を用いて行われる医師研究者主導型の研究，更にはとくにこのような支援無しで行われる自主臨床試験が含まれる．

さて治験あるいは製薬会社スポンサーの市販後臨床試験に関しては，GCP（Good Clinical Practice）あるいはこれに準じたGPMSP（Good Post Marketing Surveillance）としてその倫理性と科学性とを保証する仕組みができ上がっている．これは日・米・欧の三極の規制当局と製薬業界との国際合意ICH-GCP（ICHは「国際ハーモナイゼーション」の略称）を受け，わが国では1998年から完全施行されている．GCPとはヒトを対象とする臨床試験の計画・実施・記録および報告に対する国際的な倫理的科学的品質基準であり，この遵守によって被験者の権利・安全性および福祉が世界医師会による「ヘルシンキ宣言」に基づく原則に沿った形で保護され，臨床試験データは信頼できるという公的な保証が与えられる．このGCPに応えるため，少なくとも製薬会社主導の試験に関しては，上記のインフラストラクチャのかなりの部分がわが国でもようやく整備された．CRCの主要施設への配置やこれを派遣するSMO（Site Management Organization）の設立，施設側の体制整備，モニタリングや監査による品質保証の定着，製薬会社の体制整備，そしてメディアを通じた患者リクルートや教育も開始された．生物統計学の認知もその現れであり，製薬会社やCROには修士レベル以上の統計実務家が配置されるようになった．しかし，医師研究者主導の市販後の臨床試験に対するインフラストラクチャはまだまだ未整備である．標準治療に直結するevidenceは実はこのような市販後の臨床試験によって生み出されることが多い．わが国でevidenceの生れなかった事実は，生物統計学を含むこのような試験実施のインフラストラクチャが未整備であった結果に他ならない．

なぜ統計学が不要であったかは，わが国医学界の臨床研究軽視と根は同一である．戦後渡米した日本の俊英医学者は，言語・免許の壁から基礎医学研究の基盤を日本に導入した．これは現在花を咲かせ，一方，臨床医学の研究基盤は形成されなかった．これまでの新薬開発・薬事制度がこれに拍車をかけた．類似薬を先例通りに開発すれば認可され，保険制度の中で利益を確保できたから．わが国の臨床試験の歴史は長いが，残念ながら最近までの統計の応用はお作法であり，科学研究・技術評価の道具では必ずしもなかった．

3. 医薬品開発・標準治療確立のために行われる臨床試験

医薬品開発あるいは標準治療開発のために行われる臨床試験の流れを表1に示した．

健常人に対して行われ安全性を検討するのが第I相，患者に対して行われ，漸増的な有効性の検討がなされるのが前期第II相（海外ではIIaと呼ばれる）である．この段階では，用量に関してランダム化したいくつかの群を設定する場合も，群別にあるいは個人内で用量を漸増するデザインの双方が存在する．試験薬の安全性が十分検討されていない段階であるので，後述する「盲検」（どの薬剤を用いているのか区別できなくする工夫）無しで試験が行われることもある．対象者数を増やし，有効性と安全性のバランスから臨床用量の選択を行うのが後期第II相（IIb）である．通常は，有効成分を含まないプラセボを含んだ3から5群が設

表1 臨床試験の進め方．

	通常の薬剤	抗がん剤
第I相	●健常人対象（専門施設） ●単回，逐投試験 ●安全性検討，薬物動態	●通常は患者対象 ●複数スケジュール，増量 ●MTD・推奨用量決定
第II相	●患者対象，通常並行群 ●有効性の検討，用法用量の決定 ●安全性の検討	●患者対象，通常単群 ●腫瘍縮小による有効性の確認 ●安全性（毒性）の検討
第III相	●患者対象，通常並行群 ●標準治療との非劣性（優越性）か ●プラセボとの優越性の検証	●日本では市販後 ●標準治療との比較（通常組合わせで）

MTD: Maximum tolerated dose

定され盲検下で試験が実施される．次の第 III 相では，他に有効な対照薬が存在しない場合にはプラセボ対照，広く使われている対照薬が存在する場合にはその薬剤を対照として，当該試験薬の有効性が検証される．副作用（国際標準では有害薬物反応）対策の点で患者のリスクが過大とならない限り，盲検下で試験が実施されるのが普通である．プラセボを対照とした場合には，試験薬の優越性を検証するように，すなわち群間差の検定が有意となる確率（検出力）を一定以上にするよう試験が設計される．通常，検出力は 80–90%に設定される．実薬が対照である場合には，この優越性か，試験薬がある程度以上劣ってはいないことを検証する「非劣性試験」が設計される．わが国の治験では，それぞれの相で対象となる患者数は，おおよそ，第 I 相で 20–30 人，前期第 II 相で 30–100 人，後期第 II 相で 100–400 人，第 III 相で 200–500 人程度である．

臨床試験において有効性を測定する指標・変数がエンドポイントである．例えば降圧剤開発のエンドポイントは血圧（の降下作用），高脂血症なら総コレステロールか LDL コレステロール（の降下作用）がエンドポイントであり，これらを指標として新薬が開発されている．しかしこれらは，患者の立場に立った場合には，必ずしも生存や QOL に直結するとは限らない．降圧剤であれば，最終的な目標は血圧のコントロールを通じて脳卒中に代表される循環器系疾患を予防することである．抗高脂血症薬も主な目標は冠動脈疾患の予防である．これら最終的なエンドポイントを対象として，標準治療を確立するために行われる臨床試験は mortality・morbidity 試験と呼ばれるが，対象とした疾患の発症頻度が必ずしも高くないため，また多くのリスク因子がその発症を修飾し，後述の言葉で言えば誤差的バラツキが大きくなるため，数 1000 例ときには 10000 例を超える症例数を必要とする巨大研究となるのが普通である．当然，試験の計画から解析までには数年から 10 年を費す．有望な薬剤をできるだけ早く患者の手元に届けるため，また化合物の特許を保護し製薬会社に新薬開発のインセンティブを与える観点から，このような mortality・morbidity 試験は市販後に，しばしば製薬会社とは独立に医師研究者主導型で行い，これに先立つ認可はこれらの最終的なエンドポイントと疫学的に高い相関を有するエンドポイント評価でなされることがある．この意味で，最終的なエンドポイントに替わるエンドポイントを代替エンドポイント（surrogate endpoint）と呼ぶ．抗癌剤開発における腫瘍縮小も代替エンドポイントである（毒性が強く健常人への投与が行えない抗癌剤等では，臨床試験の組み立てがかなり異なっているがここでは省略しよう）．

表 2　最近 10 年の臨床試験の歴史．

1993.9	ソリブジン事件
1996.5	ICH-GCP（E6 ガイドライン）国際合意
1996.6	薬事法改定
1997.3	答申 GCP，省令 GCP 通知
1997.4	医薬品機構誕生
1997.7	医薬品・医療機器審査センター誕生
1998.4	GCP 完全実施
1998.8	ICH-E5 ガイドライン「海外臨床データ受け入れにおける人種要因差」通知
1998.11	ICH-E9 ガイドライン「臨床試験のための統計的原則」通知
2002.7	薬事法改定
2003.6	「医師主導の治験の実施の規準」通知 「臨床研究に関する倫理指針」通知

さて，わが国の臨床試験を取り巻く環境と実施状況は，表 2 に示すように最近 10 年間で大変革を迎えた．その引き金は前述の国際ハーモナイゼーション（ICH）であり，1993 年 9 月から 11 月の間に 15 人が副作用で亡くなったソリブジン事件であった．ヘルペスに対して開発された治療薬ソリブジンが（頻用されていたが，当時承認されていたのはわが国のみの）5FU 系統の経口抗癌剤の代謝を阻害し，重篤な薬剤相互作用を引き起こしかねないことは基礎的検討から事前に認識されていた．しかも治験中には死亡例も発生していた．しかしながら，これらの事実は軽視され，認可とそれに引き続く活発な宣伝活動により一気にソリブジンが使われ犠牲者を頻出させた．治験の質と審査体制の問題が明らかとなった本事件が主なきっかけとなり審査体制が見直され，新設の医薬品審査センターに，生物統計家 2 人が初めての統計専門審査官として厚生労働省内部に迎えられた．

製薬会社主導であれ，医師研究者主導であれ，ICH-E9 ガイドライン「統計的原則」によって臨床試験に携わる試験統計家には「適切な資格と経験（appropriately qualified and experienced）」が要求されるに至った．試験統計家の地位はわが国でも確立し，その要件と基本的な考え方も国際標準化されたといってよい．民族差に関するガイドライン（E5 ガイドライン）により，海外データを新薬申請に公式に用いることも可能となった．現在の状況は，既存の海外データを利用して国内

の治験をいかにして「サボるか」から，国際的な同時承認を目指して臨床試験をいかに国際展開するか，日本人に対する標準的な治療をいかに確立するか，再生医療や遺伝子治療などの最先端医療の臨床試験をいかに迅速に行うか，DNP チップなど遺伝子情報も生かし個別化を目指した医療（テーラーメード医療）の開発研究をいかに展開するかに移りつつある．臨床統計・生物統計の input はますます必要とされる状況が現出しつつある．

4. 臨床試験の目標と方法論の貢献

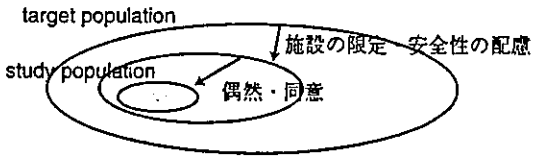
臨床試験計画時の目標は，統計的には次の 3 点にまとめられる．

1. **Clarity**：ランダム誤差を小さくし，研究の精度を向上させる
2. **Comparability**：治療群間の比較可能性（研究の正確性）を向上させる
3. **Generalizability**：結論の一般化可能性を向上させる

現状の我々の知識水準では発生理由を同定できず（あるいはその必要はなく），ランダムとみなさざるをえない（あるいはランダムとみなして操作上は差し支えない）バラツキをランダム誤差と呼び，これを小さくすることが clarity の要求である．このためのもっとも強力な対策はサンプルサイズを増やすことであり，あるいは誤差の小さな反応変数をエンドポイントとすることである．予後に大きな影響を及ぼす予後因子を計画時から考慮し（ランダム化を層化して行う，あるいは後述の動的割付法の変数に用いる），解析時に層別あるいは数学モデルにより調整することによっても研究の精度は向上する．測定やデータ管理における品質管理も重要な手段である．

Comparability の要求とは，系統誤差すなわちバイアスを減らすことに対応し，研究の正確性を向上させることと言い替えてもよい．もっとも強力な手段はランダム化であり，既知の因子については，層化要因あるいは動的割付法の変数として考慮する（当然解析時には，精度を上げるため調整の対象とすべきである）ことも重要な手段である．介入手段（通常は薬剤）の種類を医師・患者双方に対し識別不能とする 2 重盲検法は患者選択と管理・評価にわたって comparability を高く保つための手段である．なお，2 重盲検試験では，データ管理や解析の担当者も盲検化されており，この意味で 3 重盲検と呼ぶほうがふさわしい．

ランダム化 randomization　　内的妥当性の保証
無作為抽出 random sampling　外的妥当性の保証



無作為抽出：多くの統計理論において理論上重要な概念であるが，臨床試験においてはほとんど現実味をもたない．私見であるが，不用で誤解を招く概念．
Senn, S(1997), *Statistical Issues in Drug Development*, Wiley: glossary.

図 2　ランダム化と無作為抽出．

リストの最後にある generalizability は，しばしば混同して用いられる無作為抽出（random sampling）とランダム割付け（random allocation）を対比させることにより理解がしやすくなる．図 2 に示すように，臨床研究では，研究成果を適用する目標集団（target population）と実際に研究が行われる対象集団（study population）の間には数の上でも質の上でも大きな隔たりが生じる可能性が高い．まず，安全性の考慮から患者除外条件が設定され参加施設の選択がこれに続く．（例えば試験参加医師の外来担当日にたまたま来院したなど）ランダムに対象が選ばれる側面も多少はありうるが，研究に同意し参加する患者はランダム標本とは考えられない．すなわち，臨床試験研究においては（世論調査や社会調査などで統計手法の妥当性の根拠となる）無作為標本やこれを保証する無作為抽出は存在しない．この，偏りを含み得る対象集団から妥当な結論を引き出すための方法がランダム化であり，この妥当性を internal validity と呼ぶ．比較試験結果の検定に際しては，無作為標本であることは必ずしも前提として必要ではなく，ランダム化によっても，検定の妥当性は保証される（なお，臨床試験においては，E9 の翻訳において無作為化でなく「ランダム化」を標準用語として採択した．これは，後述するように単純な無作為化ではなく，多様な患者背景のバラツキを減ずるように，ランダムな要素を伴って割付けが行われる実態を強調するためである）．

したがって，厳密に言えば，得られた結論の適用範囲は当該の対象患者のみであり，この結論が目標集団まで外挿できるかどうかが一般化可能性あるいは外的

妥当性（external validity）の議論である．対象集団の背景分布が目標集団のそれと著しく異なる場合には一般化可能性は低いだろう．患者を性・年齢・疾患重症度などの背景要因によって層別し，これによって得られる部分集団（層）を解析し治療効果の差が層毎に異ならないなら，あるいは多施設共同試験では，施設内の治療効果差が施設毎に大きく異ならないなら，一般化可能性は高いだろう．アメリカの薬品・食品の認可当局 FDA (Food Drug Administration) は，第 III 相試験を最低 2 つ行うことを原則として義務付けている．これも一般化可能性を高めるためであり，メタアナリシスも一般化可能性を検討する手段である．

5. バイアスの制御とランダム化

偏り（バイアス）とは何らかの原因により，結論が系統的に真実からずれる現象であり，臨床研究においては以下の 3 つが主な原因とされている．

(1) 対象の選択
(2) 試験の実施（情報収集と評価）
(3) 交絡（Confounding）

交絡とは，患者反応に影響を及ぼす因子の分布が群間でアンバランスとなり，介入効果が正しく反応に反映しない現象である．重症の患者に薬剤が選択的に投与される結果，薬剤投与群が非投与群に劣る現象がその例である．臨床試験においては，上記 (2) に対しては盲検化によって，(1) と (3) とくに未知の因子による交絡に対してはランダム化による防止策が採用される．そして，臨床医学に対する統計学の最大の貢献はこのランダム化の導入であり，技術評価として 1930 年代に R.A. Fisher によって提唱された実験計画法とされている．新種や肥料を評価する農事試験を対象として Fisher が創案した実験計画法の基本的な考え方は，以下の概念にまとめられる．

1. バラツキの大きさを評価し，同時に推論の感度を向上させるための反復
2. 系統的なバラツキを小さくするための局所管理
3. 局所管理によっても除去しえない潜在的な，あるいは未知の要因によるバラツキを偶然誤差に転化し，確率論に基づく評価を可能とする，言い替えれば平均的に均質で比較可能な群を作るためのランダム化
4. 多数の因子の影響を効率的に評価できるよう，それらを同時に変化させて割付けを行う多因子要因実験

Fisher の実験計画法は，1947 年に英国 Medical Research Council によって開始された結核患者に対するストレプトマイシンの評価に採用され，その成功を通じ，臨床試験においても方法論としての有効性が確立されたと考えられている．臨床試験において，1 はそのまま採用された．2 は農事試験においては圃場をブロックという区画に分割することによって達成されるが，臨床試験においては重要な予後因子や参加施設で患者を事前層別することによって行われる．同一の患者に異なる治療を行う（経口薬の場合であったら日を変えて，塗布薬であったら場所を変えて）クロスオーヴァー試験は，患者個人をブロックとした研究計画である．3 は，説明と理解の上での同意という農事試験には存在しない重大な倫理的問題を引き起こしたが，臨床試験にも採用され，4 は，癌や循環器の大規模な予防研究等において積極的に採用されるに至っている．

6. 臨床試験に対する生物統計家の貢献

臨床試験に携わる生物統計家の役割の第一は，これまでの知識・先行研究を受け，臨床家との討議を通じ，データ解析によって証明できる形に仮説を設定することにある．このためには何をもって試験治療を有効とするか，いわゆる研究のエンドポイントを設定することが前提となる．新しい分野であれば，このエンドポイントの信頼性・妥当性を事前に評価することも必要となる．ランダム化試験の場合に割付けの方法を検討すること，仮説を証明するために必要最小な症例数を設定すること，偏りのない，かつ感度の高い解析法を提案すること，より探索的な相においてはデータを適切に要約するモデルを検討することも統計家の役割であり，参加患者の利益を考慮し，不必要に試験を継続しないような中間解析の方法も事前に検討される．実際のデータ解析・報告書作成に協力することも，もちろん統計家の役割である．以下，統計家の課題をやや詳しく述べよう．

(1) 証明すべき仮説は明確に，データ解析によって検証できる形に述べられているか．

検定によって新しい治療法が有意に優れることを示

す場合が多いだろうが，毒性が強い実施が困難な標準治療に対しては，試験治療の存在意義を非劣性（著しく劣ってはいない）の形で示せば十分な場合もある．この場合には閾値の設定と試験感度（assay sensitivity：本当に有効な治療が「有効」と結論されるかどうか）の確保が問題となる．後者は試験の品質管理に依存するところ大である．

現在，非劣性の検証は，臨床的に同等とみなせる差を試験治療に「上乗せ」して「明らかに劣る」という仮説を棄却し「同等以上」であることを示す検定に準じた方法と，試験治療と対照の効果の差（あるいは比）の信頼区間を計算し，これを上記の許容できる範囲と比較する方法とが用いられている．

(2) サンプルサイズの設定の根拠と方法は妥当か？

実は「必要サンプルサイズ」はきわめて曖昧で，おおよその目安を与えるにすぎない．結局は，当該分野である程度のコンセンサスを得て「臨床的に意味のある差」を設定し計算するのが実態である．十分なサンプルサイズが確保できないことが明らかならば，（必要なら海外との共同研究の形で）メタアナリシスを初めから考慮した計画とすることも考えられる．また，十分な事前情報が存在しない場合には，盲検化でこれまでのデータを中間解析し，必要サンプルサイズあるいは（長期追跡研究の場合であれば）追跡期間を延長することも行われる．最近は，独立データモニタリング委員会が開鍵した解析を行い，その結果に基づき症例数追加・追跡期間延長勧告を行う試験デザインの妥当性が議論されている．

(3) 割付け方法は妥当か？

盲検化が実施困難ながんの多施設共同研究においては，ファクシミリや電話さらには Web を用いた中央登録法はすでに常識化されている．盲検化が可能な（治験の）場合であっても，試験実施状況を迅速に把握し予後要因の偏りを防ぐ意味で，中央登録が行われる例が増えてきた．

割付けにおいて層別するとすればどのような要因を用いるか，それまでの患者の登録状況（背景要因の分布）に応じ偏りがより小さくなる方向に割付ける可能性を増やす「動的割付け法」を用いる場合には，どの因子を用いるか，施設をどう考慮するか，などが問題となる．層別要因はせいぜい 1–2，動的割付けの場合なら施設以外に 2–5 程度の因子を考慮することが可能である．施設に関しては，予見性を少なくし，かつ治療群の偏りを防ぐために動的割付けの因子とするか，極端な偏りが生じないような調整法（偏りの大きさに応じて割付確率を変える「偏コイン法」など）が用いられている．

(4) 中止・脱落・治療不遵守（違反例）例の取扱いは妥当か？

研究の目的をよく理解し，また取扱いによって選択バイアスがどのように生じるかを予測すれば，症例の取扱いは多くの場合ほぼ自明である．例えば経口剤を用いた予防研究の場合であったら，いわゆる intent-to-treat の方針に従って，有害薬物反応や患者の（途中での）投薬拒否例は分母から除くべきではないことは明らかである．有害薬物反応あるいは飲みにくさ自体が薬の負の効果であり，患者選択によって予後不良の可能性の高い患者が除かれるからである．一方，治療開始後の不適格判明については，その判定が割付けられた治療群や，治療結果に依存していないことが示されれば，解析から除くことも当然ありうる．現在では，有効性と安全性について解析対象集団を事前にプロトコルに定義するのが普通である．また，対象の取扱いによって結果がどのように変化するか「感度解析」を行うことを明記するプロトコルも増えている．

(5) 解析計画は妥当であるか，また解析結果の解釈は適切か？

検定のみならず，治療効果の推定を行うべきであることが近年強調されている．もちろん，その前提として，治療効果を単一のパラメータに縮約できるかどうかの検討が必要であり，推定結果の信頼性も信頼区間という形で示さねばならない．なお，予後に影響しない因子については，それが群間で偏りがあったとしても，調整に用いる必要はない．逆に予後に強く影響する因子については偏りが存在しなくとも解析感度（精度）を上げる意味で調整に用いるべきであり，このことはプロトコルに事前に明記することが望まれる．調整の方法としては，層別と結果の併合（連続反応変数なら分散分析，0–1 あるいは順序反応変数あるいは Time-to-event なら Cochran-Mantenl-Haenszel 流のアプローチ）あるいは回帰分析型の統計モデル（連続反応変数なら重回帰分析，0–1 反応変数ならロジスティック回帰分析，time-to-event であったら Cox 回帰）を用いるという 2 つのアプローチが用いられる．前者はラン

ダム化に確率計算の基本を置く design-based な解析法であり，後者は説明変数と反応変数の間の関連，および反応変数のバラツキに確率モデルを想定することから，model-based な方法と呼ばれている．

繰り返し測定データに対する Generalized Estimating Equations，混合効果モデル，並べ替え検定など高度・コンピュータインテンシブな手法の利用も，統計パッケージの普及もあってはもはや常識的である．

(6) 中間解析

予想以上の効果が早期から試験治療群に見られた場合，あるいは逆にこれ以上試験を継続しても試験治療の優越性（あるいは非劣性）を証明できないことが強く予想された場合に，試験を打ち切るための統計的な基準・方法が種々考案されている．代表的な方法が，事前に設定した回数の検定を多重性を考慮し行う群逐次検定（group sequential method）である．しかし，統計的基準以上に重要なことは，迅速に解析に必要なデータを収集・集計する統計センターを含むデータ管理体制であり，また，研究者とは独立な立場で評価を行うことのできる独立データモニタリング委員会（効果・安全性評価委員会）の存在である．データ固定の方法，試験途中での固定に適した調査票の設計も実務上は重要である．

臨床試験の統計的側面に関する参考文献

臨床試験の計画と統計解析に要求される水準は，近年きわめて高くなりつつあり，その状況を見るには最初の 2 つのガイドラインが適切であろう．1992 年のわが国の統計解析ガイドラインは，2) の国際合意 (ICH) ガイドラインを受け廃止された．臨床試験の統計的側面について全般については 3, 4) が，大規模試験の実施については 5) が，がん臨床試験については 7, 8) が優れた教科書である．6) は入門として価値がある．最近の動向を知るには 9) が参考となろう．サンプルサイズの設定については 10) がまとまっている．11) から 15) は最近の教科書で，11) はがんを中心とし倫理面の記述が多いことが特徴．13) は新薬開発を対象としているが，皮肉とウイットの効いた著述スタイルが面白い（日本語訳は困難）．12) はアメリカ最大のがん共同研究機構 SWOG の統計センターのスタッフによるコンパクトな医師向けのテキスト．14) は臨床薬理全般の教科書であるが，医薬品開発と統計的側面にも詳しい．16) は専門家向けの包括的な参考書（辞典）である．17) は，わが国のがん共同研究機構である JCOG のプロトコルマニュアルであり，実務上きわめて詳細かつ有用である．

1) 薬審第 335 号（1996 年 5 月 1 日）:「治験の総括報告書の構成と内容に関するガイドライン（E3)」.
2) 薬審第 1047 号（1998 年 11 月 30 日）:「臨床試験のための統計的原則（E9)」.
3) Pocock（著）・コントローラ委員会訳:『クリニカル・トライアル』，篠原出版，1989.
4) Friedman, LM., Furberg, C. and DeMets, D.: *Fundamentals of Clinical Trials*, Third ed. Springer, 1998.
5) Meinert, C.: *Clinical Trials*, Oxford Univ. Press, 1986.
6) 折笠秀樹:『臨床研究デザイン』，真興交易医書出版社，1994.
7) Leventhal, B. and Wittes, R.（著）・福島雅典・大橋靖雄（監訳）:『がん臨床研究の方法』，メディカルブックサービス，1995.
8) Buyse, M. et al. ed.: *Cancer Clinical Trials*, Oxford Univ. Press., 1985.
9) Thall, P.: *Recent Advances in Clinical Trial: Design and Analysis*, Kluwer Academic Publications, 1995.
10) Machin, D. and Campbell, M.: *Statistical Tables for the Design of Clinical Trials*, Blackwell Scientific Publications, 1987.
11) Piantadosi, S.: *Clinical Trials, A Methodologic Perspectives*, Wiley, 1997.
12) Green, S., Benedetti, J. and Crowley, J.: *Clinical Trials in Oncology* second edition, Chapman and Hall, 2003.
13) Senn, S.: *Statistical Issues in Drug Development*, Wiley, 1997.
14) 日本臨床薬理学会編:『臨床薬理学 第二版』，医学書院，2003（とくに I 章 E. 医薬品の開発）.
15) 丹後俊郎:『無作為化比較試験』，朝倉書店，2003.
16) Redmond, C. and Colton, T. ed.: *Biostatistics in Clinical Trials*, Wiley, 2001. (Wiley Reference Series in Biostatistics).
17) JCOG プロトコルマニュアル
http://jcogweb.res.ncc.go.jp/

（おおはし・やすお，東京大学大学院医学系研究科，NPO 法人・日本臨床研究支援ユニット理事長）

日本メディカルライター協会（JMCA）第3回総会・講演会　－演題1－

日本の医療情報伝達分野における問題点

日本メディカルライター協会代表理事
東京大学医学系研究科生物統計学/疫学・予防保健学
NPO日本臨床研究支援ユニット理事長
大橋　靖雄

はじめに

保健，医療に関する情報は膨大であり，正確な情報を入手するのは難しいことである。情報のなかには正確でないものも多く，伝え方がよくない場合もある。また，受け取る側にそれだけの力がないため誤って受けとめられ，健康や医療の水準を悪い方向におとしめる可能性もあるだろう。逆にいうとコミュニケーションを円滑にし，効率的にすることによって，医療や保健の水準を上げることもあるかもしれないのである。効果のない抗がん剤を開発するよりも，そのほうがどれだけ国民の健康に寄与するかしれない。

EBMには，evidenceを作り，評価・伝達し，利用するというステップがある。evidenceは臨床研究や疫学研究から生まれるが，われわれはそれを伝える努力をこれからしていかなければならないのである。

医学研究情報の氾濫

Index Medicusは1800年代末からある抄録誌である。東大図書館では現在のように情報検索にPubMed，Medlineを使用する前には，このIndex Medicusが並んでいる書庫が入り口のところにあった。あるとき，Index Medicusの重さを量ってみたところ，抄録誌が1年間で50 kgを超えていた。現在，その情報量はこの何倍にもなっており，コンピュータを使わなければ見ることができなくなっている。この膨大な情報のなかから必要な情報をいかに見分けるかという観点から，Critical Readingとか，Clinical Epidemiologyという，医師が情報をどう利用するかを教授する分野が出てきたのである。しかし，これが予防保健活動となると国民全体が利用することになるにもかかわらず，国民への教育はなく，国民を教育する立場の専門家の教育もされていない。

EBMのブームで，evidenceを作るということについては，医療関係者間にも理解が浸透しているといえるが，受け手がどう理解するかというevidenceを評価・伝達する視点が欠けている。どのような媒体に，どう発信するのか，つまり量と頻度と表現方法である。また，どの段階で発信するのかというのが医療では非常にクリティカルで難しい問題である。これらの問題にまだ答えはないが，いくつかの事例を紹介し，問題提起をしてみたい。

Physicians' Health Studyでのロイター報道事件

日本ではあまり知られていないが，1988年にPhysicians' Health Studyという大規模臨床試験にからんで起こったロイター報道事件について紹介したい。

Physicians' Health Studyはアスピリンが心血管死を減少させるか，またβカロチンが癌の発生を減少させるかを検討するために，無作為割付け，プラセボ対照，二重盲検，2×2 factorialデザインで行われ，2万2千人の医師が参加した試験である。1988年1月27日のニューヨークタイムズ等の主要紙に，心筋梗塞リスクがアスピリンを摂ることによって半減する

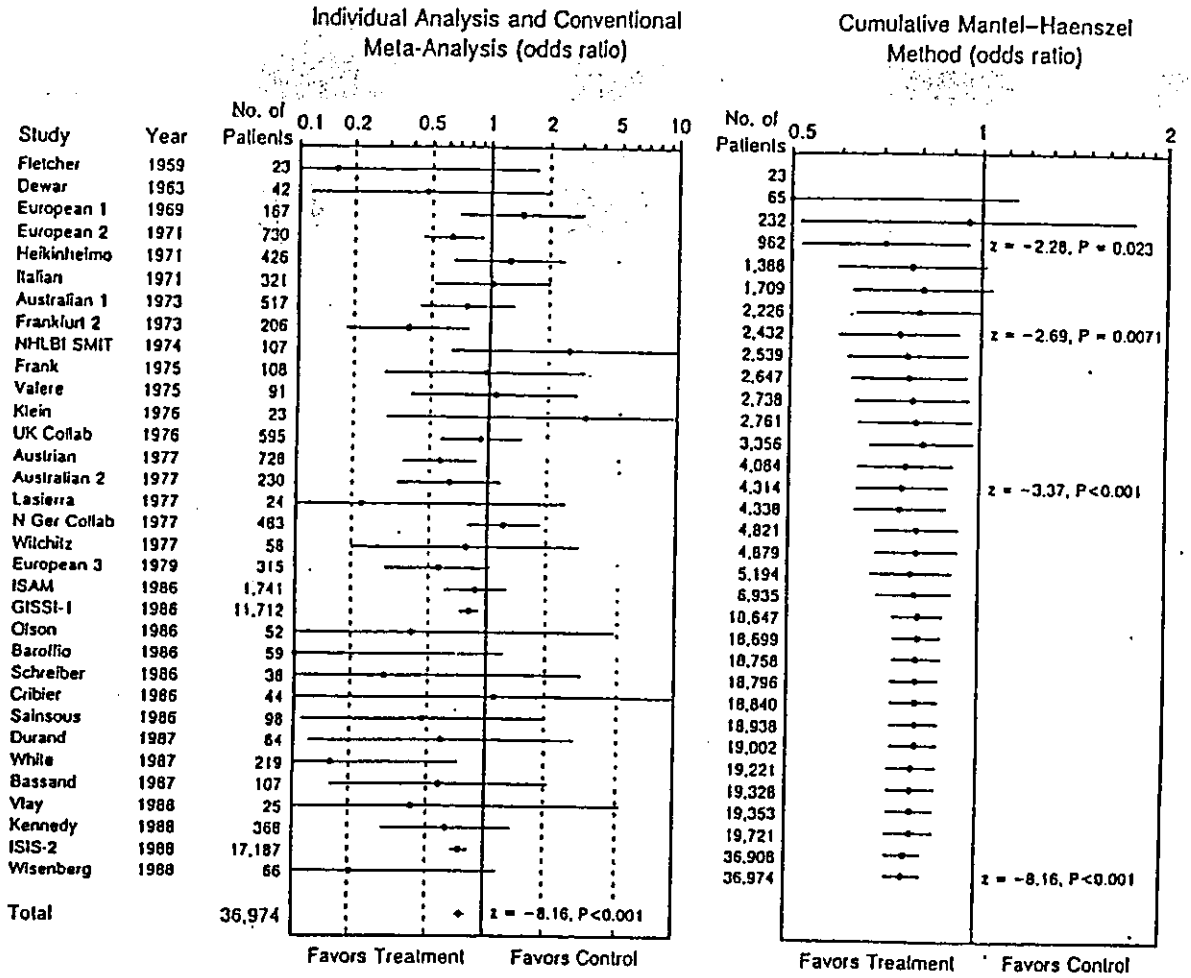


図1
(Lau J, et al. N Engl J Med. 1992 Jul 23; 327(4): 248-54 より)

という記事が掲載された。

この結果は，「Ingelfingerの原則」に従いNew England Journal of Medicine (NEJM) に報告された (N Engl J Med 1989; 321:129)。国民に対する影響が大きいため，1988年12月18日に中止勧告が出た後，緊急報告として論文が書かれ，2週間で審査終了，3週間で印刷され，1月28日に発行されたが，このような迅速な対応はおそらくこれまでなかっただろう。

「Ingelfingerの原則」とは，NEJMが1982年に経験した苦い事件を教訓にして，編集者であるIngelfingerが作った原則である。これは糖尿病患者を対象とした大規模臨床試験であるthe University Group Diabetes Program (UGDP) の結果から，経口の血糖降下剤を飲むことにより，心臓疾患死が増えるというデータが論文発表前にマスコミに流れてしまったため，医師のもとに，「私が飲んでいる薬はあれではないか」という患者が殺到しパニックになったという事件である。この時，医師は情報をまったく持っていなかったため，その後NEJMは，研究者や医療者に情報が十分行き渡るまでは，一般への情報公開は禁止するとして，逆にそれを守らなければ論文を載せないという方針を作ったのである。そして，この教訓に従い1月21日に印刷論文250部をマスコミに送付し，1週間後の28日を解禁日としたが，ロイターが解禁日前にこれを報道してしまった。これに対し，NEJMは情報送付停止という報復措置をとったが，この報復に対して，試験に参加していた医師の間から「私はこの治験に参加していたが，もっと早く情報を提供すべきではなかったか。NEJMのやり方はあまりにも硬直的ではないか。12月18日の段階で