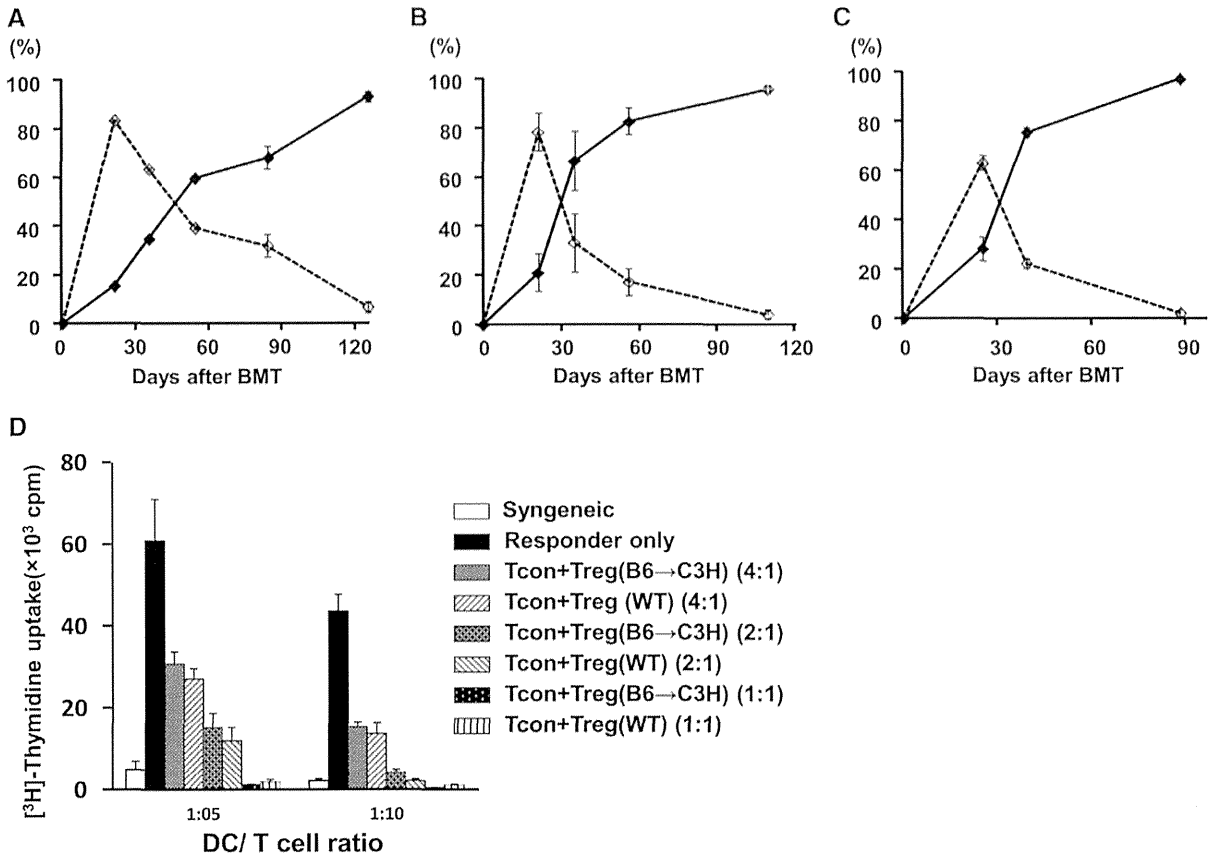


**Figure 2.** Donor BM–derived progenitors comprise the long-term peripheral Treg pool. Lethally irradiated C3H recipients underwent transplantation as in Figure 1: (B6 → C3H). The rates of CD45.2$^+$ spleen cell–derived (broken lines) and CD45.2$^-$ BM–derived (solid lines) Treg in CD4$^+$Foxp3$^+$ Treg are shown. Spleen (A) and mesenteric lymph nodes (MLN) (B) were isolated from (B6 → C3H) mice at various time points after BMT and cells were analyzed by fluorescent activated cell sorter. (C) Lethally irradiated C3.SW (H-2$^b$) recipients underwent transplantation from B6 (H-2$^b$) donors. The rates of CD45.2$^+$ splenic T cell–derived (broken lines) and CD45.2$^-$ BM–derived (solid lines) Treg in CD4$^+$Foxp3$^+$ Treg in the spleen are shown. Each group consisted of 20 to 23 mice. The means (±SE) of each group are shown. Data are from a representative of at least 2 independent experiments. (D) CD25$^+$CD4$^+$ Treg were purified from the spleens of (B6 → C3H) mice (on day 120) or naïve B6 (WT). B6 CD4$^+$CD25$^-$ T cells (Tcon) together with various numbers of Treg were cultured with irradiated C3H CD11c$^+$ DC as stimulators for 72 hours. Proliferative activities were determined by monitoring $^3$H-thymidine uptake.

parameters: weight loss, posture, activity, fur texture, and skin integrity (maximum index, 10), as described previously [22]. Shaved skin from the interscapular region (approximately 2 cm$^2$), liver, and salivary gland specimens of recipients were fixed in 10% formalin, embedded in paraffin, sectioned, mounted on slides, and stained with hematoxylin and eosin. Skin slides were scored on the basis of dermal fibrosis, fat loss, inflammation, epidermal interface changes, and follicular drop out (0 to 2 for each category; the maximum score was 10) [21]. Liver slides were scored based on bile duct injury and inflammation (0 to 4 for each category), and the maximum score was 8 [25]. Salivary gland slides were scored based on atrophy and inflammation (0 to 3 for each category), and the maximum score was 6. All slides were scored by pathologists (T.K. and T.T.) blind to experimental group.

#### *Immunohistochemistry*

Immunohistochemical staining for Foxp3 and CD3 was performed using the high polymer (HISTOFINE simple stain, NICHIREI, Tokyo, Japan) method. Anti-Foxp3 (eBioscience) and anti-CD3 (Abcam, Cambridge, MA) were used to identify Tregs and effector T cells, respectively.

#### *Flow Cytometry*

The mAbs used were unconjugated anti-CD16/32 (2.4G2); FITC-, PE-, PerCP-, or APC-conjugated anti-mouse CD4, CD25, CD45.1, CD45.2, H-2$^b$, H-2$^d$ (BD Pharmingen, San Diego, CA); and Foxp3 (eBioscience, San Diego, CA), as described previously [26]. A Foxp3 staining kit (eBioscience) was used for intracellular staining. Cells were analyzed on a FACSAria flow cytometer with FACSDiva software (BD Immunocytometry Systems, San Diego, CA).

#### *Mixed Leukocyte Reaction*

CD4$^+$CD25$^-$ T cells, CD4$^+$CD25$^+$ T cells, and CD11c$^+$ DC were magnetically separated by AutoMACS using microbeads from a CD4$^+$CD25$^+$ regulatory T cell isolation kit and CD11c microbeads. CD4$^+$CD25$^-$ T cells ($5 \times 10^4$ per well) together with various numbers of CD25$^+$CD4$^+$ T cells (0 to $5 \times 10^4$ per well) were cultured with irradiated (30 Gy) CD11c$^+$ DC as stimulators for 72 hours in 96-well round-bottomed plates. Cells were pulsed with $^3$H-thymidine (1 μCi [.037 MBq] per well) for a further 16 hours [27]. Proliferation was determined using Topcount NXT (Packard Instruments, Meriden, CT).

#### *Statistics*

Data are given as means ± SEM. The survival curves were plotted using Kaplan-Meier estimates. Group comparisons of pathology scores were performed using the Mann-Whitney *U* test. Comparative analysis of cell ratios was performed by the unpaired 2-tailed Student *t*-test or Welch's *t*-test. In all analyses, $P < .05$ was taken to indicate statistical significance.

## RESULTS

### *Kinetics of Treg Reconstitution after Allogeneic BMT*

We first examined whether Tregs intermixed in the graft persist in the host for long periods post BMT using the MHC-mismatched model of BMT. Lethally irradiated C3H (H-2$^k$) recipient mice received $10 \times 10^6$ TCD-BM cells from B6.Ly-5a (H-2$^b$,CD45.1) mice with/without 1 to $2 \times 10^6$ spleen cells from B6 (H-2$^b$,CD45.2) mice. All of the recipients of allogeneic C3H TCD-BM cells from B6 mice and syngeneic mice survived and were resistant to induction of GVHD. Although 100% of

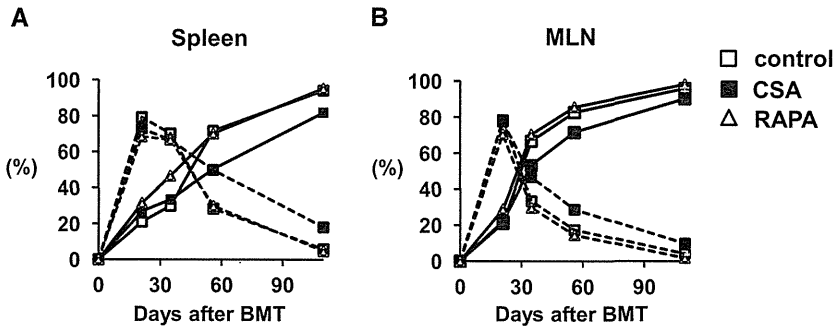


**Figure 3.** Effects of CSA and mTOR inhibitors on the Treg compartment. Lethally irradiated C3H recipients underwent transplantation from B6 donor mice as shown in Figure 1 and received i.p. injections of CSA (closed squares), mTOR inhibitor (rapamycin, RAPA; open triangles), or vehicle control (open squares) daily from day 0 to 110. The rates of $CD45.2^+$ splenic T cell–derived (broken lines) and $CD45.2^-$ BM–derived (solid lines) Treg in $CD4^+Foxp3^+$ Treg are shown. Spleen (A) and mesenteric lymph nodes (MLN) (B) were isolated from (B6 → C3H) mice at various time points after BMT and cells were analyzed by fluorescent activated cell sorter. Each group consisted of 16 to 23 mice. The means (±SE) of each group are shown. Data are from a representative of at least 2 independent experiments.

the animals that received allogeneic BM plus $2 \times 10^6$ spleen cells died by day 35 with clinical and histopathological signs of severe GVHD, the recipients of allogeneic BM plus $1 \times 10^6$ spleen cells (BM plus Sp cells) showed mild clinical signs of GVHD and 60% survived by day 120 (Figure 1A); the following experiment was performed in this setting. Flow cytometric analysis of donor cell chimerism in the spleen 3 weeks after allogeneic BMT showed that 98.8% ± 0.7% of spleen cells were derived from the donor in mice, thus confirming complete donor cell engraftment. Host Tregs, as determined by $CD4^+Foxp3^+H\text{-}2^{k+}$, were not detected in the spleen on day 21 post transplantation (data not shown). On day 21 post transplantation, the majority of $CD4^+Foxp3^+$ Tregs were derived from $CD45.2^+$ splenic T cells (83.4% ± 2.2%), suggesting that splenic T cell–derived Tregs underwent homeostatic and/or alloantigen-driven expansion (Figure 1B) and the absolute number of Tregs in the spleens of the recipients of BM plus Sp cells was significantly higher than in TCD-BM recipients. From day 21 onward, due to GVHD-induced lymphopenia, the absolute number of Tregs in the spleens of recipients of BM plus Sp cells was lower than in TCD-BM recipients (Figure 1C). The rate of $CD45.2^+$ splenic T cell–derived Tregs in $CD4^+Foxp3^+$ Treg decreased gradually and most $CD4^+Foxp3^+$ Treg were $CD45.1^+$ BM-derived (93.2%) on day 125 post transplantation (Figure 2A). The rate of $CD45.1^+$ BM-derived Tregs in the mesenteric lymph nodes (MLN) was also increased and became dominant in the late post-transplantation period (Figure 2B). To exclude strain-dependent artifacts, we next evaluated the kinetics of Treg reconstitution in the B6 ($H\text{-}2^b$) into C3.SW ($H\text{-}2^b$) MHC-compatible, multiple minor histocompatibility antigen (miHA)-incompatible model of SCT. The kinetics of Treg reconstitution in the spleen was similar and most $CD4^+Foxp3^+$ Tregs were derived from $CD45.1^+$ BM (97%) on day 90 post transplantation (Figure 2C). These findings indicated that the peripheral Treg pool was restored first by expanded splenic T cell–derived mature Treg and then by new Tregs generated from donor BM-derived progenitors. Next, to examine the function of newly arising Tregs, purified $CD4^+CD25^+$ T cells on day 120 post transplantation were

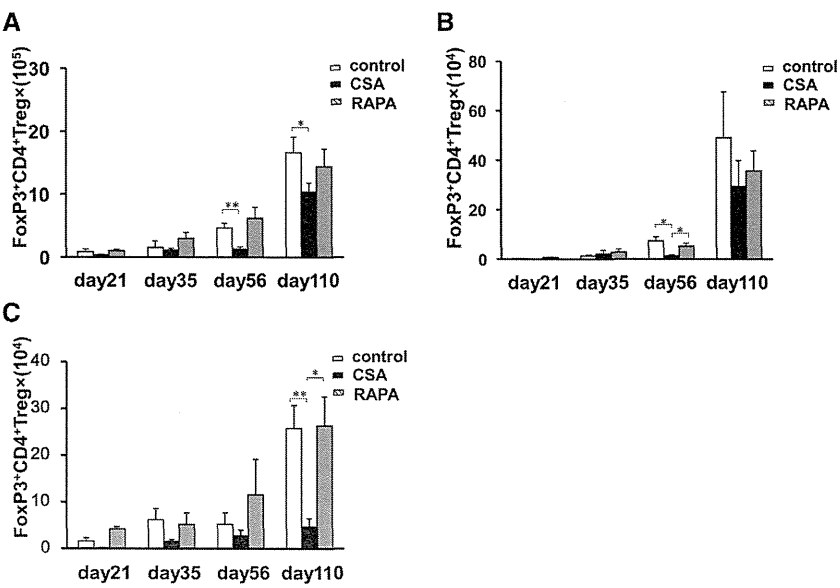


**Figure 4.** CSA, but not mTOR, inhibitors hampered reconstitution of BM-derived Treg. (B6 → C3H) mice received i.p. injections of CSA (black bars), mTOR inhibitor (rapamycin, RAPA; gray bars), or vehicle control (white bars) daily from day 0 to 110. The absolute numbers of Treg in the spleen (A), MLN (B), and thymus (C) are shown. Each group consisted of 19 to 26 mice. The means (±SE) of each group are shown. Data are from a representative of at least 2 independent experiments. $^*P < .05$; $^{**}P < .01$.

assessed for their ability to inhibit proliferation by responding syngeneic $CD4^+CD25^-$ B6 T cells. Their suppressive activity was virtually indistinguishable from that of Tregs obtained from normal B6 mice (Figure 2D). Taken together, Tregs generated from donor BM-derived progenitors comprise the long-term peripheral Treg pool and exhibit immunosuppressive activity.

### *CSA, but Not mTORinhibitors, Hampered Reconstitution of BM-derived Treg*

Coenen et al. reported that 28 days of CSA administration hampered Treg homeostasis in normal mice [28]. We examined whether the use of CSA for an extended period affected the long-term peripheral Treg pool after BMT. C3H recipient mice underwent transplantation from B6 donor mice (as shown in Figure 1) and received i.p. injection of CSA, mTOR inhibitor (rapamycin; RAPA), or vehicle control daily from day 0. We analyzed the effects of CSA and RAPA on the Treg compartment at 21, 35, 56, and 110 days post hematopoietic cell transplantation. Mice treated with CSA or RAPA showed the same Treg reconstitution pattern as those treated with vehicle solution. On day 21 post transplantation, the majority of $CD4^+Foxp3^+$ Tregs in the spleen were $CD45.2^+$ splenic T cell–derived cells but the Treg compartments were dominated by BM-derived cells on days 56 and 110 post transplantation in all 3 groups (Figure 3A). In the MLN, these 3 groups also showed similar Treg reconstitution kinetics (Figure 3B). There were no differences in the absolute numbers of Treg among the 3 groups on day 21. From day 21 onward, however, the absolute numbers of Tregs in the CSA-treated mice were lower than those in control mice both in the spleen (day 56: $1.3 \pm .4$ versus $4.6 \pm .8 \times 10^5$, $P < .01$; day 110: $10.4 \pm 1.4$ versus $16.7 \pm 2.4 \times 10^5$, $P < .05$) (Figure 4A) and in the MLN (day 56: $1.3 \pm .5$ versus $7.4 \pm 1.6 \times 10^4$, $P < .03$; day 110: $2.9 \pm 1.0$ versus $4.9 \pm 1.9 \times 10^5$, $P = .46$) (Figure 4B). Especially in the thymus, mice treated with CSA showed a marked reduction in the absolute numbers of Tregs compared with those treated with vehicle control (day 110: $4.6 \pm 1.8$ versus $25.7 \pm 5.0 \times 10^4$, $P < .01$) (Figure 4C). In contrast to mice treated with CSA, mice treated with RAPA showed no reduction in the absolute numbers of Tregs and no differences compared with control mice in the spleen or MLN at any time point post transplantation (Figure 4A,B). The absolute numbers of newly arising Tregs in the thymus were also not reduced in mice treated with RAPA (Figure 4C). We next examined the effects of another mTOR inhibitor, everolimus (RAD), which exhibits greater polarity than RAPA and has been approved in Europe for use as an immunosuppressant for prevention of cardiac and renal allograft rejection. Reconstitution of newly arising Tregs in the thymus was not impaired in mice treated with RAD, and there were no differences in the absolute numbers of spleen Tregs compared with control mice on day 110 (spleen: $15.4 \pm 2.5$ versus $16.6 \pm 2.4 \times 10^5$, $P = .73$, Supplemental Figure 1A; thymus: $17.4 \pm 3.2$ versus $25.7 \pm 5.0 \times 10^4$, $P = .26$, Supplemental Figure 1B). These findings suggested that CSA, but not mTOR inhibitors, hampered the long-term reconstitution of BM-derived Tregs.

### *CSA, but Not mTOR Inhibitors, Increased Liability to Chronic GVHD*

Recent studies revealed the association of reduced Treg frequency in patients with chronic GVHD. In the present study, we examined histopathological change in CSA-treated mice where reconstitution of BM-derived Tregs was impaired. The skin of CSA-treated mice showed pathogenic features of chronic GVHD (Figure 5A), and pathogenic scores revealed significantly exacerbated chronic GVHD pathology compared with those treated with vehicle control ($5.5 \pm .8$ versus $1.6 \pm .3$, $P < .01$) (Figure 5B). A dry mouth is one of the distinctive features of chronic GVHD. Lymphocytic inflammation, fibrosis, and atrophy of acinar tissue were observed in the salivary glands of CSA-treated mice (Figure 5A) and pathological scores were significantly higher in CSA-treated

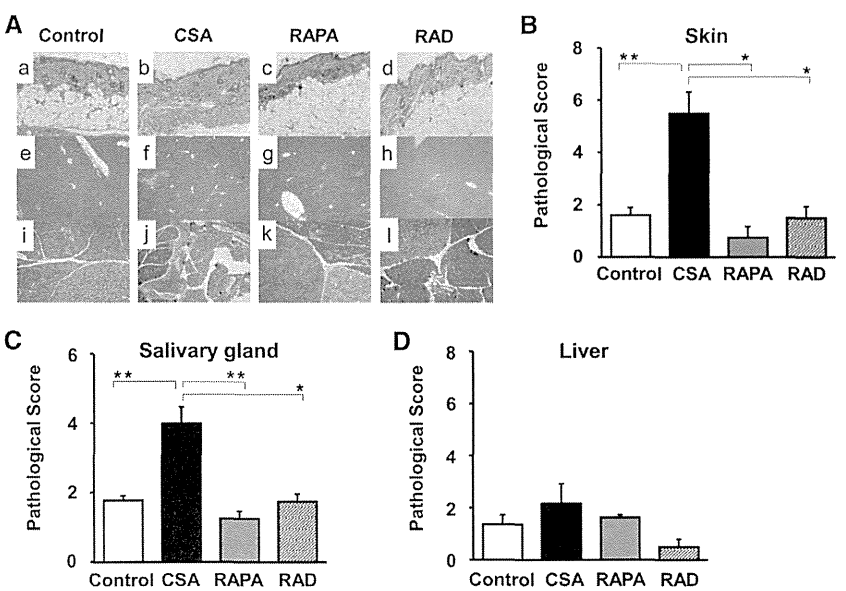


**Figure 5.** CSA, but not mTOR, inhibitors increased the likelihood of chronic GVHD. (A) Histological findings of the skin (a to d), liver (e to h), and salivary glands (i to l) (on day 120) from (B6 → C3H) mice given CSA, mTOR inhibitor (RAPA, RAD), or vehicle control. Sclerodermatous skin changes, such as epidermal atrophy, fat loss, follicular dropout, and dermal thickness (b); fibrosis in the portal area and peripheral mononuclear cells infiltrates in the liver (f); and fibrosis and atrophy of acinar tissue in the salivary glands (j) were observed (original magnification: × 100) Pathological scores of skin (B), salivary gland (C) and liver (D). The data are expressed as means ± SE. Data are from a representative of at least 2 independent experiments. *$P < .05$; **$P < .01$.

mice than in the controls (4.0 ± .5 versus 1.8 ± .1, $P < .01$) (Figure 5C). CSA-treated mice showed bile duct injury and fibrosis in the portal area and peripheral mononuclear cell infiltration in the liver and pathological scores of the liver also tended to be worse in CSA-treated mice, as compared with those treated with vehicle control, although it was not statistically significant (Figure 5D). In contrast to mice treated with CSA, mice treated with RAPA showed no pathogenic features of chronic GVHD and there were no differences in pathogenic skin and salivary gland scores, as compared with control mice (skin: .75 ± .4 versus 1.6 ± .3, $P = .18$, Figure 5B; salivary gland: 1.25 ± .2 versus1.78 ± .1, $P = .08$, Figure 5C). Immunohistochemical staining for Foxp3 and CD3 revealed that $CD3^+$ T cells infiltrated in the skin tissue of all 3 groups, and RAD-treated mice showed abundant infiltration by $CD3^+$ T cells and $Foxp3^+$ cells (Figure 6A). In contrast to RAD, $Foxp3^+$ cells were scarcely found in skin tissue of CSA-treated mice. The ratio of Foxp3 Tregs per 100 $CD3^+$ lymphocytes in the skin tissue of CSA-treated mice was significantly lower than those in RAD-treated mice (3.23 ± .4 versus 19.5 ± 4.4, $P < .05$). CSA-treated mice tended to show poorer survival, as compared with those treated with mTOR inhibitors or vehicle control (CSA 27.6% versus control 54.2%, RAD 57.1%, RAPA 61.5%, $P = .28$, Supplemental data Figure 2). These findings suggested that CSA, but not mTOR inhibitors, hampered the reconstitution of BM-derived Treg and increased liability to chronic GVHD.

We next tested liability to chronic GVHD in CSA-treated mice using adoptive transfer experiments. Previously, Sakoda et al. demonstrated that impaired thymic negative selection of the recipients permitted the emergence of pathogenic T cells that cause chronic GVHD (Figure 7A) [23]. Lethally irradiated C3H recipients were reconstituted with TCD BM from MHC class II-deficient ($H2\text{-}Ab1^{-/-}$) B6 mice ([$H2\text{-}Ab1^{-/-} \rightarrow$C3H]). These mice developed disease conditions that showed all of the clinical and histopathological features of human chronic GVHD. $CD4^+$ T cells isolated from chronic GVHD mice ([$H2\text{-}Ab1^{-/-} \rightarrow$C3H] $CD4^+$ T cells) cause chronic GVHD when B6 antigens are provided by hematopoietic cells in the absence of B6 antigen expression on target epithelium ([B6→C3H] chimeras) [23]. In the current study, C3H mice underwent transplantation from B6 donors as shown in Figure 1 and were orally administered CSA, RAPA, or vehicle solution until 60 days post BMT, when none of the recipients showed significant signs of chronic GVHD. To test liability to chronic GVHD, these C3H-recipient mice with B6-derived antigen presenting cells received adoptive transfer of [$H2\text{-}Ab1^{-/-} \rightarrow$C3H] $CD4^+$ T cells (Figure 7B). As shown in Figure 7C and D, adoptive transfer of pathogenic $CD4^+$ T cells caused severe weight loss (CSA 81.1 ± 4.1% versus control 94.5 ± 2.1%, $P < .05$; and CSA 81.1 ± 4.1% versus RAPA 98.9 ± 1.5%, $P < .01$) and chronic GVHD in CSA-treated mice, with a mortality rate of 83%. RAPA-treated mice and controls showed resistance to induction of chronic GVHD by transfer of pathogenic $CD4^+$ T cells; the

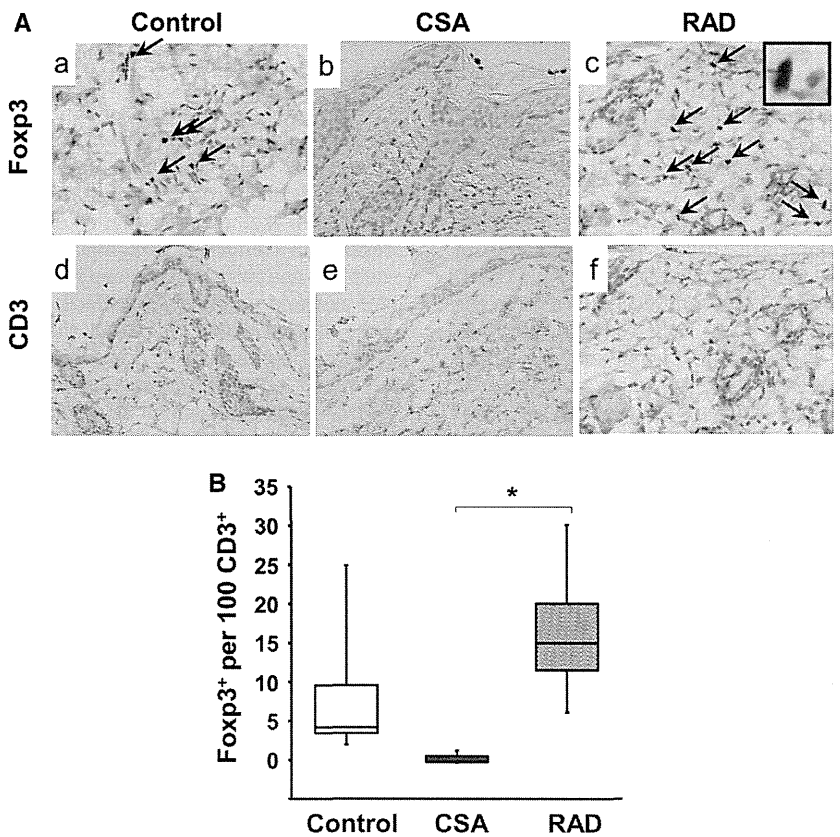


**Figure 6.** CSA, but not mTOR, reduces Treg infiltration in skin tissue. (A) Lethally irradiated C3H recipients underwent transplantation from B6 donor mice as shown in Figure 1 and received vehicle control (a, d), CSA (b, e), or mTOR inhibitor (RAD; c, f), daily from day 0 to 120. Immunohistochemical staining was performed using anti-Foxp3 (a to c) and anti-CD3 (d to f) antibodies on day 120. Arrows indicate Foxp3 positive cells. (B) The ratio of Foxp3 Tregs per 100 $CD3^+$ lymphocytes. The number of CD3 and Foxp3 cells was counted in all the high-power fields. Results are expressed as mean ± SD. Pictures and data are from a representative of 2 independent experiments. (n = 3 to 4 per group). *$P < .05$.

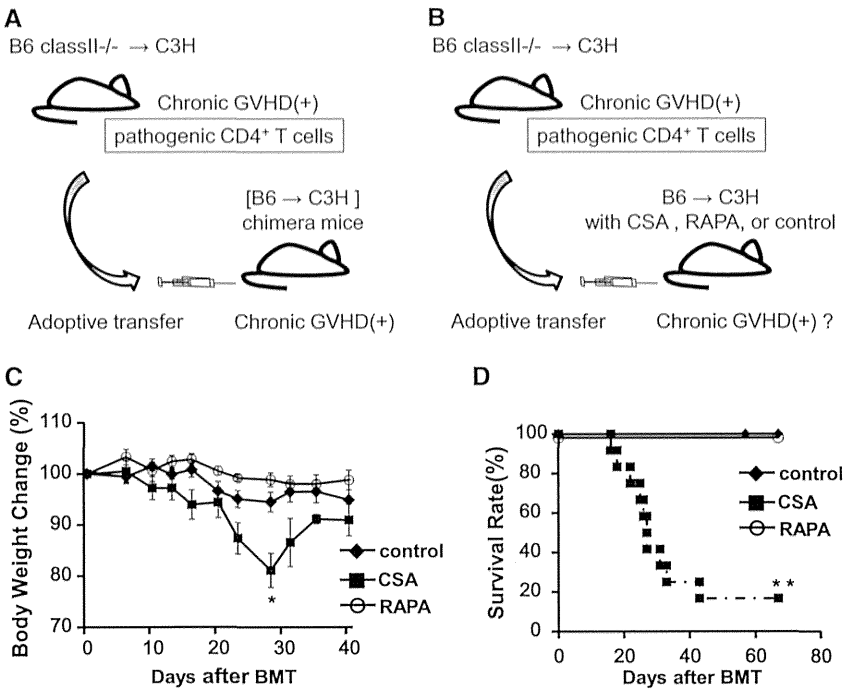


**Figure 7.** Adoptive transfer of pathogenic $CD4^+$ T cells caused severe chronic GVHD. (A) Lethally irradiated C3H recipients were reconstituted with TCD BM from MHC class II-deficient ($H2\text{-}Ab1^{-/-}$) B6 mice. These mice developed chronic GVHD and $CD4^+$ T cells isolated from chronic GVHD mice ([$H2\text{-}Ab1^{-/-}$→C3H] $CD4^+$ T cells) were primarily donor reactive. These pathogenic $CD4^+$ T cells cause chronic GVHD when B6 antigens are provided by hematopoietic cells in the absence of B6 antigen expression on target epithelium ([B6→C3H] chimeras). (B) C3H recipient mice underwent transplantation from B6 donors as shown in Figure 1 and received CSA, RAPA, or vehicle solution until 60 days post BMT. These C3H recipient mice received adoptive transfer of [$H2\text{-}Ab1^{-/-}$→C3H] $CD4^+$ T cells. Body weight change is shown in (C) and overall survival is shown in (D). Data from 3 similar experiments are combined (n = 8 to 12 per group). The data are expressed as means ± SE. $^*P < .05$; $^{**}P < .01$.

survival rate on day 62 after adoptive transfer was 100%. Taken together, these data demonstrated that CSA, but not mTOR inhibitors, increased liability to chronic GVHD.

## DISCUSSION

Patients with chronic GVHD have a lower frequency of Tregs when compared with patients without chronic GVHD [29-32]. Experimental BMT demonstrated that Tregs in the inoculum can prevent acute GVHD when injected together with donor T cells [12-14]; however, it is not known whether Tregs in the grafts persist into the late post-transplantation period and play a role in preventing chronic GVHD. Mastuoka et al. prospectively monitored $CD4^+$ T cell subsets and showed that thymic generation of naïve Treg was markedly impaired and Treg levels subsequently declined in patients with prolonged $CD4^+$ lymphopenia [32]. This resulted in a relative Treg deficiency, which was associated with a high incidence of extensive chronic GVHD. In the present study, we monitored Treg reconstitution kinetics in the spleen, MLN, and thymus according to 2 subsets, T cells derived from peripheral-expanded mature T cells and newly arising T cells from bone marrow stem cells, using 2 mouse BMT models because this is difficult to examine in a human setting. The results indicated that host Tregs disappeared rapidly in mice receiving allogeneic T cells early in the early post-transplantation period, consistent with a previous report [33]. In addition, this study showed that splenic T cell–derived Treg initially occupy a niche in lymphopenic transplantation recipients, suggesting that mature Treg underwent homeostatic and/or alloantigen-driven expansion. However, the donor splenic T cell–derived Treg pool contracted gradually and Tregs generated from donor BM-derived progenitors comprised the long-term peripheral Treg pool. The BM-derived Treg compartment was functionally competent, as determined by in vitro lymphoid suppression, indicating that these cells play a role in post-BMT immune tolerance.

Coenen et al. reported that 28 days of treatment with CSA resulted in a reduction in thymic generation of $CD4^+Foxp3^+$ T cells and peripheral $CD25^+Foxp3^+$ T cells in normal mice [28]. We assessed whether CSA affects the peripheral Treg pool after allogeneic BMT; on day 21, there were no differences in the absolute numbers of Tregs among 3 groups, and CSA had no impact on early Treg reconstitution. Consistent with our observations, Setoguchi et al. reported that in contrast to the requirement of IL-2 for physiological expansion of $CD4^+CD25^+$ Treg cells in normal nonlymphopenic mice, homeostatic proliferation in a lymphopenic environment appears to be IL-2-independent [19]. Zeiser et al. also reported that CSA administration has only a minor impact on the expansion of adoptively transferred $CD4^+CD25^+$ T cells on day 7 post transplantation [34]. However, whether prolonged use of CSA affects the long-term peripheral Treg pool has not been reported. Our data showed that CSA, but not mTOR inhibitors, hampered the long-term reconstitution of BM-derived Tregs. The numbers of Tregs in the spleen, thymus and tissue were significantly reduced in mice receiving CSA in comparison with those receiving mTOR inhibitors or PBS on day 110. CSA blocks nuclear factor of activated T cells translocation into the nucleus by inhibiting calcineurin phosphatase activity [35]. CSA inhibits the thymic generation of Tregs by impairment of TCR signaling and by reducing nuclear factor of activated T cells–dependent Foxp3 promoter activity [36]. In contrast, rapamycin-sensitive downstream targets of phosphatidyl-inositol 3-kinase are IL-2-independent, and

rapamycin affects neither the initial signal transduction upon TCR triggering nor the thymic generation of Treg [37]. Immunosuppressive drugs have different mechanisms of promoting immune suppression and our data revealed different effects on the long-term peripheral Treg pool after allogeneic BMT.

Although mouse models of chronic GVHD have provided important insights into pathophysiology of this disease, one factor that confounds the translation of findings in mouse models to the human disease is that time course of development of chronic GVHD is more rapid in most mouse models than in human. Another factor is that most patients are given immunosuppressive therapy to prevent acute GVHD [38], and these medications might influence the development of chronic GVHD. In this study, histopathological examination revealed that CSA-treated mice showed pathogenic features of chronic GVHD, whereas those treated with mTOR inhibitors showed no significant differences compared with control mice. This is the first report that long-term use of CSA induces chronic GVHD in transplant-recipient mice. This may have been due to induction of autoreactive T cells by CSA [39,40]. Wu et al. reported that CSA contributes to chronic GVHD in experimental models, which was ascribed to the disruption of clonal-deletion mechanisms in the thymus, resulting in the export of autoreactive T cells [41]. The present study demonstrated another mechanism by which CSA impaired Treg reconstitution. Adoptive transfer of the pathogenic $CD4^+$ T cells caused severe chronic GVHD in CSA-treated mice, whereas mTOR inhibitor-treated and control mice showed resistance to induction of chronic GVHD. These findings suggest that the increased liability to chronic GVHD in CSA-treated mice might be due to limited reconstitution of BM-derived Treg cells; further mechanistic studies will be required to determine if this is truly causative rather than merely an association.

Here, we assessed the differential effects of CSA and mTOR inhibitors on the long-term peripheral Treg pool after allogeneic BMT. Our findings indicated that, in contrast to mTOR inhibitors, CSA compromises homeostasis in peripheral immune compartments and thymic generation of $CD4^+CD25^+Foxp3^+$ T cells. GVHD prophylaxis with mTOR inhibitor and calcineurin inhibitor failed to reduce chronic GVHD [11,42-45]. The choice of calcineurin inhibitor–free GVHD prophylaxis with mTOR inhibitors, such as mTOR inhibitors plus IL-2 [16] or mTOR inhibitors plus antithymocyte globulin [46] may have important implications for the control of chronic GVHD after BMT.

## ACKNOWLEDGMENTS

*Financial disclosure:* The authors have nothing to disclose.

*Conflict of interest statement:* There are no conflicts of interest to report.

## SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.bbmt.2013.11.018.

## REFERENCES

1. Teshima T, Wynn TA, Soiffer RJ, et al. Chronic graft-versus-host disease: how can we release Prometheus? *Biol Blood Marrow Transplant.* 2008; 14:142-150.
2. Socie G, Stone JV, Wingard JR, et al. Long-term survival and late deaths after allogeneic bone marrow transplantation. Late Effects Working Committee of the International Bone Marrow Transplant Registry. *The New England journal of medicine.* 1999;341:14-21.
3. Baker KS, Gurney JG, Ness KK, et al. Late effects in survivors of chronic myeloid leukemia treated with hematopoietic cell transplantation: results from the Bone Marrow Transplant Survivor Study. *Blood.* 2004; 104:1898-1906.
4. Deeg HJ, Lin D, Leisenring W, et al. Cyclosporine of cyclosporine plus methylprednisolone for prophylaxis of graft-versus-host disease: a prospective, randomized trial. *Blood.* 1997;89:3880-3887.
5. Storb R, Deeg HJ, Pepe M, et al. Graft-versus-host disease prevention by methotrexate combined with cyclosporin compared to methotrexate alone in patients given marrow grafts for severe aplastic anaemia: long-term follow-up of a controlled trial. *Br J Haematol.* 1989;72:567-572.
6. Kansu E, Gooley T, Flowers ME, et al. Administration of cyclosporine for 24 months compared with 6 months for prevention of chronic graft-versus-host disease: a prospective randomized clinical trial. *Blood.* 2001;98:3868-3870.
7. Couriel DR, Saliba R, Escalon MP, et al. Sirolimus in combination with tacrolimus and corticosteroids for the treatment of resistant chronic graft-versus-host disease. *Br J Haematol.* 2005;130:409-417.
8. Johnston LJ, Brown J, Shizuru JA, et al. Rapamycin (sirolimus) for treatment of chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2005;11:47-55.
9. Jurado M, Vallejo C, Perez-Simon JA, et al. Sirolimus as part of immunosuppressive therapy for refractory chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2007;13:701-706.
10. Jedlickova Z, Burlakova I, Bug G, et al. Therapy of sclerodermatous chronic graft-versus-host disease with mammalian target of rapamycin inhibitors. *Biol Blood Marrow Transplant.* 2011;17:657-663.
11. **Cutler C, Logan BR, Nakamura R, et al. Tacrolimus/sirolimus vs. tacrolimus/methotrexate for graft-vs.-host disease prophylaxis after hla-matched, related donor hematopoietic stem cell transplantation: results of Blood and Marrow Transplant Clinical Trials Network Trial 0402. [ASH Annual Meeting Abstracts] Blood 2012;120:739.**
12. Edinger M, Hoffmann P, Ermann J, et al. CD4+CD25+ regulatory T cells preserve graft-versus-tumor activity while inhibiting graft-versus-host disease after bone marrow transplantation. *Nature Med.* 2003;9:1144-1150.
13. Hoffmann P, Ermann J, Edinger M, et al. Donor-type CD4(+)CD25(+) regulatory T cells suppress lethal acute graft-versus-host disease after allogeneic bone marrow transplantation. *J Exp Med.* 2002;196:389-399.
14. Taylor PA, Lees CJ, Blazar BR. The infusion of ex vivo activated and expanded CD4(+)CD25(+) immune regulatory cells inhibits graft-versus-host disease lethality. *Blood.* 2002;99:3493-3499.
15. Zeiser R, Leveson-Gower DB, Zambricki EA, et al. Differential impact of mammalian target of rapamycin inhibition on CD4+CD25+Foxp3+ regulatory T cells compared with conventional CD4+ T cells. *Blood.* 2008;111:453-462.
16. Shin HJ, Baker J, Leveson-Gower DB, et al. Rapamycin and IL-2 reduce lethal acute graft-versus-host disease associated with increased expansion of donor type CD4+CD25+Foxp3+ regulatory T cells. *Blood.* 2011;118:2342-2350.
17. Fontenot JD, Rasmussen JP, Gavin MA, Rudensky AY. A function for interleukin 2 in Foxp3-expressing regulatory T cells. *Nature Immunol.* 2005;6:1142-1151.
18. D'Cruz LM, Klein L. Development and function of agonist-induced CD25+Foxp3+ regulatory T cells in the absence of interleukin 2 signaling. *Nature Immunol.* 2005;6:1152-1159.
19. Setoguchi R, Hori S, Takahashi T, Sakaguchi S. Homeostatic maintenance of natural Foxp3(+) CD25(+) CD4(+) regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization. *J Exp Med.* 2005;201:723-735.
20. Grusby MJ, Johnson RS, Papaioannou VE, Glimcher LH. Depletion of CD4+ T cells in major histocompatibility complex class II-deficient mice. *Science.* 1991;253:1417-1420.
21. Anderson BE, McNiff JM, Matte C, et al. Recipient CD4+ T cells that survive irradiation regulate chronic graft-versus-host disease. *Blood.* 2004;104:1565-1573.
22. Reddy P, Maeda Y, Liu C, et al. A crucial role for antigen-presenting cells and alloantigen expression in graft-versus-leukemia responses. *Nature Med.* 2005;11:1244-1249.
23. Sakoda Y, Hashimoto D, Asakura S, et al. Donor-derived thymic-dependent T cells cause chronic graft-versus-host disease. *Blood.* 2007;109:1756-1764.
24. Matsumoto Y, Hof A, Baumlin Y, et al. Differential effects of everolimus and cyclosporine A on intimal alpha-actin-positive cell dynamics of carotid allografts in mice. *Transplantation.* 2004;78:345-351.
25. Kaplan DH, Anderson BE, McNiff JM, et al. Target antigens determine graft-versus-host disease phenotype. *J Immunol.* 2004;173:5467-5475.
26. Duffner UA, Maeda Y, Cooke KR, et al. Host dendritic cells alone are sufficient to initiate acute graft-versus-host disease. *J Immunol.* 2004; 172:7393-7398.
27. Maeda Y, Tawara I, Teshima T, et al. Lymphopenia-induced proliferation of donor T cells reduces their capacity for causing acute graft-versus-host disease. *Exp Hematol.* 2007;35:274-286.
28. Coenen JJ, Koenen HJ, van Rijssen E, et al. Rapamycin, not cyclosporine, permits thymic generation and peripheral preservation of CD4+ CD25+ FoxP3+ T cells. *Bone Marrow Transplant.* 2007;39:537-545.

29. Miura Y, Thoburn CJ, Bright EC, et al. Association of Foxp3 regulatory gene expression with graft-versus-host disease. *Blood.* 2004;104:2187-2193.
30. Rieger K, Loddenkemper C, Maul J, et al. Mucosal FOXP3+ regulatory T cells are numerically deficient in acute and chronic GvHD. *Blood.* 2006; 107:1717-1723.
31. Zorn E, Kim HT, Lee SJ, et al. Reduced frequency of FOXP3+ CD4+CD25+ regulatory T cells in patients with chronic graft-versus-host disease. *Blood.* 2005;106:2903-2911.
32. Matsuoka K, Kim HT, McDonough S, et al. Altered regulatory T cell homeostasis in patients with CD4+ lymphopenia following allogeneic hematopoietic stem cell transplantation. *J Clin Invest.* 2010;120:1479-1493.
33. Bayer AL, Jones M, Chirinos J, et al. Host CD4+CD25+ T cells can expand and comprise a major component of the Treg compartment after experimental HCT. *Blood.* 2009;113:733-743.
34. Zeiser R, Nguyen VH, Beilhack A, et al. Inhibition of CD4+CD25+ regulatory T-cell function by calcineurin-dependent interleukin-2 production. *Blood.* 2006;108:390-399.
35. Dunn CJ, Wagstaff AJ, Perry CM, et al. Cyclosporin: an updated review of the pharmacokinetic properties, clinical efficacy and tolerability of a microemulsion-based formulation (neoral)1 in organ transplantation. *Drugs.* 2001;61:1957-2016.
36. Mantel PY, Ouaked N, Ruckert B, et al. Molecular mechanisms underlying FOXP3 induction in human T cells. *J Immunol.* 2006;176:3593-3602.
37. Abraham RT, Wiederrecht GJ. Immunopharmacology of rapamycin. *Annual review of immunology.* 1996;14:483-510.
38. Blazar BR, Taylor PA, Panoskaltsis-Mortari A, Vallera DA. Rapamycin inhibits the generation of graft-versus-host disease- and graft-versus-leukemia-causing T cells by interfering with the production of Th1 or Th1 cytotoxic cytokines. *J Immunol.* 1998;160:5355-5365.
39. Hess AD, Fischer AC, Horwitz L, et al. Characterization of peripheral autoregulatory mechanisms that prevent development of cyclosporin-induced syngeneic graft-versus-host disease. *J Immunol.* 1994;153: 400-411.
40. Bucy RP, Xu XY, Li J, Huang G. Cyclosporin A-induced autoimmune disease in mice. *J Immunol.* 1993;151:1039-1050.
41. Wu DY, Goldschneider I. Cyclosporin A-induced autologous graft-versus-host disease: a prototypical model of autoimmunity and active (dominant) tolerance coordinately induced by recent thymic emigrants. *J Immunol.* 1999;162:6926-6933.
42. Cutler C, Li S, Ho VT, et al. Extended follow-up of methotrexate-free immunosuppression using sirolimus and tacrolimus in related and unrelated donor peripheral blood stem cell transplantation. *Blood.* 2007;109:3108-3114.
43. Alyea EP, Li S, Kim HT, et al. Sirolimus, tacrolimus, and low-dose methotrexate as graft-versus-host disease prophylaxis in related and unrelated donor reduced-intensity conditioning allogeneic peripheral blood stem cell transplantation. *Biol Blood Marrow Transplant.* 2008;14: 920-926.
44. Rodriguez R, Nakamura R, Palmer JM, et al. A phase II pilot study of tacrolimus/sirolimus GVHD prophylaxis for sibling donor hematopoietic stem cell transplantation using 3 conditioning regimens. *Blood.* 2010;115:1098-1105.
45. Rosenbeck LL, Kiel PJ, Kalsekar I, et al. Prophylaxis with sirolimus and tacrolimus +/- antithymocyte globulin reduces the risk of acute graft-versus-host disease without an overall survival benefit following allogeneic stem cell transplantation. *Biol Blood Marrow Transplant.* 2011;17:916-922.
46. Schleuning M, Judith D, Jedlickova Z, et al. Calcineurin inhibitor-free GVHD prophylaxis with sirolimus, mycophenolate mofetil and ATG in Allo-SCT for leukemia patients with high relapse risk: an observational cohort study. *Bone Marrow Transplant.* 2009;43: 717-723.

*Original Article*

www.diahome.org

Therapeutic Innovation
& Regulatory Science
1-7

DOI: 10.1177/2168479014546333
tirs.sagepub.com

# Incorporating Historical Data in Bayesian Phase I Trial Design: The Caucasian-to-Asian Toxicity Tolerability Problem

**Kentaro Takeda, MS[1,2], and Satoshi Morita, PhD[1,3]**

**Abstract**

Following phase I dose-finding oncology trials completed in Western countries, Asian investigators often conduct further phase I trials to determine the maximum tolerated dose for Asian patients. This may be due to concerns about possible differences in treatment tolerability between Caucasian and Asian patient groups. Our proposed approach aims to appropriately borrow strength from a previous Caucasian trial to improve the maximum tolerated dose determination in an Asian population of patients. We design an Asian phase I trial using the Bayesian continual reassessment method. First we analyze toxicity data from a Caucasian trial to derive the prior distributions for a subsequent Asian trial. Then, we calibrate the informativeness of the prior distributions according to prior effective sample size defined by Morita et al. Extensive simulation studies demonstrate favourable operating characteristics of the proposed method, compared with two methods based on power and noninformative priors, respectively.

**Keywords**

dose finding, phase I study design, historical data, prior effective sample size, continual reassessment method

## Introduction

In this paper, we propose an approach to incorporating historical data to establish prior distributions for a dose-finding clinical trial to develop an anticancer agent. Following phase I dose-finding trials completed in Western countries, Asian investigators often conduct further phase I trials to determine the maximum tolerated dose (MTD) for Asian patients. This may be due to concerns about possible differences in treatment tolerability between Caucasian and Asian patient groups. In several cases, different treatment MTDs were estimated for Asians and Caucasians.[1,2] For example, phase I studies of capecitabine (Xeloda) monotherapy undertaken in Caucasians identified 1657 mg/m$^2$ as the MTD.[3,4] After these studies were completed, a phase I trial in Japanese patients determined a higher dose level, 2510 mg/m$^2$, as the MTD for Japanese patients.[5] Taking into account the recent trend of the globalization of new drug development, it may be worth considering the relevant use of historical data from a previous trial to design and conduct a subsequent trial in a new region. It should, however, be noted that an overly use of prior information may rather degrade the study design of a subsequent trial.

Our proposed approach aims to appropriately borrow strength from a previous Caucasian trial to improve the MTD determination in an Asian population of patients. We design an Asian phase I dose-finding trial using the Bayesian continual reassessment method,[6,7] even if other Bayesian designs can be used. The continual reassessment method is a model-based method that enables us to utilize all available prior information

Supplementary material for this article is available on the journal's website at http://tirs.sagepub.com/supplemental.

[1] Department of Biostatistics and Epidemiology, Graduate School of Medicine, Yokohama City University, Yokohama, Japan
[2] Biostatistics Group, Data Science, Global Development, Astellas Pharma Inc, Tokyo, Japan
[3] Department of Biomedical Statistics and Bioinformatics, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Submitted 05-Apr-2014; accepted 15-Jul-2014

**Corresponding Author:**
Kentaro Takeda, Biostatistics Group, Data Science, Global Development, Astellas Pharma Inc, 2-5-1, Nihonbashi-Honcho, Chuo-ku, Tokyo 103-8411, Japan.
Email: kentaro.takeda@astellas.com

through prior distributions of the model parameters. First, we analyze toxicity data from a Caucasian trial to derive the priors for a subsequent Asian trial. We suppose that the Caucasian trial is conducted using a traditional "3 + 3" cohort design[8] and that the same dose levels are tested commonly in Caucasian and Asian trials. Second, we calibrate the informativeness of the priors according to a prior effective sample size (ESS).[9,10] Morita et al[10] wrote that the prior ESS provides a useful tool for understanding the impact of prior assumptions in Bayesian study design and data analysis. We call these priors based on the prior ESS "ESS priors." Finally, we conduct the Asian phase I trial using the continual reassessment method with the ESS priors.

In our study, we compare our proposed method with two methods based on power and noninformative priors, respectively, in terms of their performance in estimating MTD in a subsequent Asian dose-finding study. The power prior was proposed by Ibrahim and Chen[11] to allow investigators to incorporate and downweight historical data. The power prior raises the likelihood of historical data to a power parameter, $a_0 \in [0, 1]$, that controls how much strength to borrow from the historical data: "no borrowing ($a_0 = 0$)" to "full borrowing ($a_0 = 1$)."

This paper is organized as follows. In the next section, we outline the Bayesian study designs of an Asian phase I trial incorporating historical data from a previously conducted Caucasian phase I trial. We conduct extensive simulation studies to examine the operating characteristics of our proposed method in the subsequent section. We close with a brief discussion.

## Probability Model and Study Designs

We compare the methods embedded with the 3 types of priors: the ESS, power, and noninformative priors. Note that the difference among the 3 methods is only in establishing the priors that are to be assumed in the Asian trial.

### *Preliminary Notation and Probability Model for Toxicity*

Let $\mathcal{D}_C$ and $\mathcal{D}_A$ denote data from the Caucasian and Asian trials, respectively. That is, $\mathcal{D}_C$ and $\mathcal{D}_A$ correspond to the historical data and the current study data, respectively. Suppose that both Caucasian and Asian phase I trials are conducted to investigate a single anticancer agent with the same dose levels. Each patient receives one of $J$ doses denoted by $d_1, \ldots, d_J$, with standardized doses $x_j = d_j / s.d.(d_1, \cdots, d_J)$, where, *s.d.* abbreviates standard deviation. As described in the introduction, we suppose that, for simplicity, the same dose levels are tested commonly in Caucasian and Asian trials. However, it is not difficult to extend our proposed method to cases where different dose levels are examined between two populations of patients.

We use a two-parameter logistic model to derive the priors based on the previous Caucasian data, as well as to design and conduct a subsequent Asian phase I trial. The outcome variable is the indicator $Y_1 = 1$ if a patient $i$ suffers toxicity, 0 if not. Denoting the probability of toxicity under dose $x_i$ by $\pi(x_i, \alpha, \beta)$, we assume the following two-parameter logistic model,

$$\pi(x_i, \alpha, \beta) = \Pr(Y_i = 1 | x_i, \alpha, \beta) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \quad (1)$$

with the intercept and slope parameters $\alpha$ and $\beta$. We assume a normal prior for $\alpha$ as

$$\alpha \sim N(\mu_\alpha, \gamma_\alpha) \quad (2)$$

To constrain $\beta$ to be positive, we assume a gamma prior for $\beta$ as

$$\beta \sim Ga(g_1(\mu_\beta, \gamma_\beta), g_2(\mu_\beta, \gamma_\beta)), \quad (3)$$

where $\mu_\beta$ and $\gamma_\beta$ are the prior mean and variance of $\beta$, respectively, and $g_1(s, t) = s^2/t$ and $g_2(s, t) = s/t$. We assume that $\alpha$ and $\beta$ are a priori independent. We use Markov chain Monte Carlo to compute the posteriors,[12] because the joint posterior distribution of regression coefficient parameters is not readily available in closed form.

### *Establishing ESS Prior*

By analyzing the historical data $\mathcal{D}_C$ using the two-parameter logistic model (1), we compute the posterior means and variances of $\alpha$ and $\beta$ that are denoted by $(\tilde{\mu}_{\alpha,C}, \tilde{\mu}_{\beta,C})$ and $(\tilde{\gamma}_{\alpha,C}, \tilde{\gamma}_{\beta,C})$, respectively. For the priors of the model parameters in the Asian phase I trial, we propose to assume

$$\begin{aligned} \alpha &\sim N(\tilde{\mu}_{\alpha,C}, w \cdot \tilde{\gamma}_{\alpha,C}), \\ \beta &\sim Ga(g_1(\tilde{\mu}_{\beta,C}, w \cdot \tilde{\gamma}_{\beta,C}), g_2(\tilde{\mu}_{\beta,C}, w \cdot \tilde{\gamma}_{\beta,C})), \end{aligned} \quad (4)$$

where $w$ is a constant for the prior calibration. Then we calibrate the prior distributions by tuning $w$ so that the priors have a prior ESS, $m$.[9,10] That is, we use the prior ESS as a guide to calibrate the priors. A prior ESS quantifies the prior information in terms of an equivalent number of hypothetical patients. As described in the next section, in the simulation study we will vary the values of $m$ (e.g, $m = 1, 2, \ldots, 10$) to examine the impact of the prior informativeness on the operating characteristics of the study design. The algorithm to derive the prior distributions is summarized as follows:

Step 1: Retrospectively analyze $\mathcal{D}_C$ to estimate $\tilde{\mu}_{\alpha,C}$, $\tilde{\mu}_{\beta,C}$, $\tilde{\gamma}_{\alpha,C}$, and $\tilde{\gamma}_{\beta,C}$ using the model ,

Step 2: Calibrate the informativeness of the priors (4) by tuning $w$ according to the prior ESS.

In step 1, we use the priors (2) and (3) to stabilize the retrospective analysis of $\mathcal{D}_C$. We obtain the two estimates of the probabilities of toxicity at two doses, the second lowest ($d_2$)

**Table 1.** True dose-toxicity relationships (true toxicity probability at 6 doses) under 2 scenarios in Caucasian patients and 6 scenarios in Asian patients.

| | Dose Level | | | | | |
|---|---|---|---|---|---|---|
| Scenario | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
| Caucasian | | | | | | |
| 1 | 0.01 | 0.05 | 0.10 | **0.30** | 0.50 | 0.60 |
| 2 | 0.01 | 0.02 | 0.03 | 0.05 | 0.10 | **0.30** |
| Asian | | | | | | |
| 1 | 0.01 | 0.05 | 0.10 | **0.30** | 0.50 | 0.60 |
| 2 | 0.05 | 0.14 | **0.36** | 0.65 | 0.86 | 0.95 |
| 3 | 0.10 | **0.30** | 0.50 | 0.60 | 0.70 | 0.80 |
| 4 | **0.41** | 0.58 | 0.82 | 0.94 | 0.98 | 0.99 |
| 5 | 0.01 | 0.02 | 0.03 | 0.05 | 0.10 | **0.30** |
| 6 | 0.03 | 0.06 | 0.12 | 0.21 | **0.36** | 0.53 |

Maximum tolerated doses are shown in boldface.

and second highest ($d_{J-1}$), from preclinical study data. These two probabilities give the prior means $\mu_\alpha$ and $\mu_\beta$.[7] Then, we assume the common prior variance for $\alpha$ and $\beta$ (ie, $\gamma_\alpha = \gamma_\beta$) that is specified as having an appropriate amount of prior information (prior ESS = 3) so that the priors never dominate the posterior inference.[9,10]

### *Power and Noninformative Priors*

In this study, we use the most basic version of power prior, that is, the power prior with a fixed $a_0 \in [0, 1]$, rather than expressing uncertainty about $a_0$ by using an additional prior distribution.[13] With $\mathcal{D}_C$ as historical data, we denote the historical likelihood by $L(\alpha, \beta|\mathcal{D}_C)$. This likelihood is specified by the two-parameter logistic model (1). We use the following conditional power prior for the parameters $\alpha$ and $\beta$ in the Asian trial,

$$p(\alpha, \beta|D_C, a_0) \propto L(\alpha, \beta|\mathcal{D}_C)^{a_0} p(\alpha, \beta). \quad (5)$$

In this paper we define $a_0$ as $a_0 = n_C/N_C$, where $N_C$ is the total number of patients treated in the previous Caucasian trial and $n_C$ is an integer $\in [1, N_C]$. Note that $n_C$ in this power prior plays the same role of $m$ in the ESS prior.[9] In the simulation study, we similarly vary the values of $n_C$ from 1 to an appropriate number $< N_C$ as with $m$. With respect to $p(\alpha, \beta)$, we assume a noninformative normal prior $N(0, 10000)$ for $\alpha$ and a noninformative gamma prior $Ga(0.0001, 0.0001)$ with mean 1 and variance 10,000 for $\beta$. We also use the same noninformative priors of $\alpha$ and $\beta$ in the third method that is based on noninformative priors.

### *Trial Conduct*

Recall that we suppose that the Caucasian phase I trial was conducted with the traditional "3 + 3" cohort design. The Caucasian trial started the dose escalation at the lowest dose $d_1$. After the MTD in the Caucasian patients was determined according to the "3 + 3" design, 12 patients were added on the MTD level as an expansion cohort.

We carry out an Asian phase I dose-finding trial using the continual reassessment method. That is, we have a continual reassessment method–type goal of finding the "optimal" dose $x^*$. Optimal is defined as the posterior mean of $\pi(x^*)$ being closest to some fixed target $\pi^*$. The maximum sample size is 30, with the cohort size of 3, starting at the lowest dose $d_1$ and not skipping a dose level when escalating, with target toxicity probability $\pi^* = 0.33$. Dose assignment is based on the posterior distribution conditional on all data available at the time of the decision. This allows for a precise estimation of the dose level with expected toxicity closest to the desired target toxicity.

## Simulation Studies

### *Simulation Study Design*

We studied the performance of the proposed study design embedded with the ESS prior (ESS design) by comparing it to the two other designs with the power and noninformative priors (power design and noninformative design) in the setting of a new phase I trial in Asian patients. As summarized in Table 1, we constructed 2 and 6 different toxicity scenarios specifying dose-toxicity relations in the Caucasian and Asian patient groups, respectively. Under all 12 combinations of the 2 and 6 scenarios, we simulated the Caucasian and Asian trials 3000 times. That is, in each of the 3000 simulations, we first generated one set of Caucasian data, analyzed the data for the prior derivation, and then simulated one subsequent Asian trial. The same basic setup for the Asian trial simulations was used in all 3 designs for a fair comparison with respect to the dose levels (= 6 levels; 100, 200, 300, 400, 500, 600 mg), the

maximum number of patients per trial (= 30), cohort size (= 3), starting dose (= $d_1$), and target $\pi^* = .33$. We investigated the operating characteristics of the ESS design under $m = 1, 3$, and 10, and those of the power design similarly under $n_C = 1, 3$, and 10, as described above. As reference, in the simulations of Caucasian trials, the average number of patients per trial was around 30. Thus, for example, $n_C = 3$ on average corresponds to $a_0 = 0.1$ in the simulations. We carried out the simulation study using the integer values $m = 1$ to 10. Although we drew the figures using all the values of $m$ from 1 to 10, we described the simulation results limited to the values of $m = 1$, 3, and 10 in the tables.

### Simulation Results

The operating characteristics for the 3 designs are summarized in Table 2, which is organized into scenario sections. The results are summarized in terms of the percentage of times that each design selected each dose level as the final MTD and the percentages of patients who were treated at each dose level. Correct selection percentages are given in boldface. We also report the average number of patients experiencing toxicity in the trial. The simulation results with the 6 Asian scenarios under Caucasian scenario 1 are shown in Table 2. For each scenario section, the first rows represent the true toxicity probabilities at the 6 dose levels in Asian population of patients.

Under Caucasian scenario 1 and Asian scenario 1, both patient groups have the same MTD (= $d_4$). The ESS and power designs more correctly selected $d_4$ as the MTD than the noninformative design, obviously due to the prior information derived from the preceding phase I trial. With increasing $m$ and $n_C$ (incorporating more prior information), the percentage of correct final recommendations gradually increased in both the ESS and power designs.

Under Asian scenario 2, the ESS and power designs more correctly selected the MTD than the noninformative design. The correct selection percentages were higher than those obtained under Asian scenario 1, even for the noninformative design. These high percentages may be in part due to the setup of the relatively high true toxicity probability $d_4$ (= .65), which may lead to decreasing the selection of $d_4$ and increasing the selection of $d_3$.

Under Asian scenario 3, the ESS design appeared to perform better than the power design in terms of selecting $d_2$ as the MTD for Asian patients. The difference in the performance between those two designs may be partly due to the formulations of the embedded priors. The power prior (5) in a sense directly incorporated toxicity data observed at each of the 6 doses. Thus, it seemed that the power design more intensely reflected the Caucasian data, especially that observed at upper dose levels (ie, $d_4$ and $d_5$) than the ESS design. In the simulations of the Caucasian trial, 28.9% and 9.3% of patients were treated at $d_4$ and $d_5$, respectively. In contrast, the ESS design, in this paper, constructed the two separate priors for the intercept and slope parameters by analyzing the preceding trial data. This formulation might lead to more desirable performance of the ESS design. In addition, and more interesting, it seemed that the ESS design might have an optimal range of prior informativeness (ie, prior ESS, $m$) that provides the best performance under several conditions of the study design. Figure 1 shows the percentages of final MTD recommendation at each dose level with respect to prior ESS values ($m = 0$ to 10) under Asian scenario 3. The correct MTD selection (= $d_2$) percentages got the highest mark in between $m = 1$ and 3, perhaps because the ESS priors with such prior informativeness played an important role as a useful guide for dose escalation/de-escalation decisions early in the trial, and after enrolling 3 patients, the information from the likelihood started to dominate the prior, as desired. The results under the other scenarios are shown in Appendix Figure S1.

Under Asian scenarios 4 and 5, even the noninformative design worked sufficiently well. As expected, the frequency of correct MTD selection gradually decreased in both the ESS and power designs as $m$ and $n_C$ went up. Under Asian scenario 6, the ESS design seemed to perform somewhat better than the power design.

Under Caucasian scenario 2, results and findings were similar to those under Caucasian scenario 1 (Appendix Table S1).

## Discussion

We have proposed an approach to quantifying prior information from a previous dose-finding trial to design a subsequent trial in a different population of patients via specified prior distributions. Our proposal is to calibrate the derived priors according to a prior ESS. It is motivated by the idea that one may avoid the use of an overly informative prior in the sense that inference is dominated by the prior rather than the data. Our simulations show that our proposed method has more advantages over the other two methods based on the power and noninformative priors in terms of their performance at estimating MTD in a subsequent Asian dose-finding study.

Several limitations to our proposed approach should be kept in mind. Our approach heavily depends on the model assumption—that is, the two-parameter logistic model for the dose-toxicity relationship. As always, the robustness of our approach to the model assumption should be evaluated before being recommended for practical use. Furthermore, the essential disadvantage of our approach may be in using the information obtained from one single previous study to derive priors for a subsequent trial in a different patient population. To deal with this issue, an extension of our method to combine multiple

**Table 2.** Simulation results using designs based on the effective sample size prior, power prior, and noninformative prior for a subsequent phase I trial in Asian patients under Caucasian scenario 1.

| | | Dose Level | | | | | | Allocation % | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | MTD | >MTD | Average Toxicity |
| Caucasian scenario 1 | True toxicity prob. | 0.01 | 0.05 | 0.10 | **0.30** | 0.50 | 0.60 | | | |
| Asian scenario 1 | True toxicity prob. | 0.01 | 0.05 | 0.10 | **0.30** | 0.50 | 0.60 | | | |
| Noninformative prior | | 0.1 | 0.5 | 9.6 | **54.4** | 27.6 | 7.9 | 34.2 | 25.2 | 7.8 |
| Effective sample size prior | | | | | | | | | | |
| $m = 1$ | | 0.0 | 0.0 | 6.3 | **63.0** | 27.3 | 3.4 | 36.6 | 26.7 | 8.1 |
| $m = 3$ | | 0.0 | 0.0 | 5.9 | **63.5** | 28.2 | 2.5 | 38.0 | 25.6 | 8.0 |
| $m = 10$ | % Recommendation | 0.0 | 0.0 | 5.0 | **67.3** | 26.4 | 1.4 | 42.8 | 22.5 | 7.9 |
| Power prior | | | | | | | | | | |
| $n_c = 1$ | | 0.2 | 0.5 | 7.9 | **58.1** | 27.4 | 5.8 | 36.5 | 26.3 | 8.1 |
| $n_c = 3$ | | 0.2 | 0.2 | 7.3 | **61.2** | 25.8 | 5.3 | 37.7 | 26.2 | 8.2 |
| $n_c = 10$ | | 0.1 | 0.3 | 8.0 | **64.5** | 22.5 | 4.6 | 40.0 | 23.6 | 8.0 |
| Asian scenario 2 | True toxicity prob. | 0.05 | 0.14 | **0.36** | 0.65 | 0.86 | 0.95 | | | |
| Noninformative prior | | 0.7 | 20.8 | **68.5** | 8.8 | 1.2 | 0.0 | 45.9 | 13.6 | 9.1 |
| Effective sample size prior | | | | | | | | | | |
| $m = 1$ | | 0.2 | 18.4 | **76.9** | 4.4 | 0.1 | 0.0 | 50.6 | 13.1 | 9.3 |
| $m = 3$ | | 0.0 | 15.8 | **79.2** | 5.0 | 0.0 | 0.0 | 53.8 | 13.2 | 9.6 |
| $m = 10$ | % Recommendation | 0.0 | 8.7 | **85.2** | 6.0 | 0.0 | 0.0 | 58.3 | 16.9 | 10.4 |
| Power prior | | | | | | | | | | |
| $n_c = 1$ | | 1.0 | 18.5 | **72.0** | 7.9 | 0.6 | 0.1 | 47.0 | 17.2 | 9.8 |
| $n_c = 3$ | | 1.8 | 12.7 | **75.9** | 9.0 | 0.5 | 0.0 | 49.2 | 19.7 | 10.3 |
| $n_c = 10$ | | 1.8 | 8.4 | **77.6** | 12.0 | 0.2 | 0.0 | 48.2 | 25.2 | 11.2 |
| Asian scenario 3 | True toxicity prob. | 0.10 | **0.30** | 0.50 | 0.60 | 0.70 | 0.80 | | | |
| Noninformative prior | | 12.3 | **56.8** | 25.3 | 4.4 | 0.9 | 0.3 | 46.4 | 29.0 | 9.5 |
| Effective sample size prior | | | | | | | | | | |
| $m = 1$ | | 5.1 | **68.9** | 24.1 | 1.7 | 0.1 | 0.0 | 51.4 | 31.6 | 10.0 |
| $m = 3$ | | 2.2 | **68.4** | 27.9 | 1.5 | 0.0 | 0.0 | 51.6 | 35.0 | 10.5 |
| $m = 10$ | % Recommendation | 0.2 | **53.4** | 44.2 | 2.2 | 0.0 | 0.0 | 36.5 | 53.3 | 11.9 |
| Power prior | | | | | | | | | | |
| $n_c = 1$ | | 9.9 | **54.5** | 30.5 | 4.2 | 0.7 | 0.2 | 40.1 | 37.8 | 10.4 |
| $n_c = 3$ | | 10.0 | **48.1** | 35.9 | 5.3 | 0.5 | 0.2 | 34.0 | 45.1 | 10.9 |
| $n_c = 10$ | | 9.7 | **33.9** | 47.8 | 8.0 | 0.5 | 0.1 | 25.4 | 57.4 | 12.2 |
| Asian scenario 4 | True toxicity prob. | **0.41** | 0.58 | 0.82 | 0.94 | 0.98 | 0.99 | | | |
| Noninformative prior | | **96.0** | 3.9 | 0.1 | 0.0 | 0.0 | 0.0 | 88.2 | 11.8 | 13.1 |
| Effective sample size prior | | | | | | | | | | |
| $m = 1$ | | **96.1** | 3.9 | 0.0 | 0.0 | 0.0 | 0.0 | 83.1 | 16.9 | 13.3 |
| $m = 3$ | | **92.7** | 7.3 | 0.0 | 0.0 | 0.0 | 0.0 | 73.6 | 26.5 | 13.8 |
| $m = 10$ | % Recommendation | **67.1** | 32.9 | 0.0 | 0.0 | 0.0 | 0.0 | 36.7 | 63.3 | 16.2 |
| Power prior | | | | | | | | | | |
| $n_c = 1$ | | **92.6** | 6.8 | 0.6 | 0.0 | 0.0 | 0.0 | 82.4 | 17.6 | 13.5 |
| $n_c = 3$ | | **92.1** | 7.2 | 0.6 | 0.0 | 0.0 | 0.0 | 78.9 | 21.1 | 13.8 |
| $n_c = 10$ | | **88.6** | 10.1 | 1.1 | 0.2 | 0.0 | 0.0 | 67.1 | 32.9 | 14.9 |
| Asian scenario 5 | True toxicity prob. | 0.01 | 0.02 | 0.03 | 0.05 | 0.10 | **0.30** | | | |
| Noninformative prior | | 0.0 | 0.0 | 0.0 | 0.4 | 9.1 | **90.4** | 40.6 | — | 4.5 |
| Effective sample size prior | | | | | | | | | | |
| $m = 1$ | | 0.0 | 0.0 | 0.0 | 0.1 | 7.6 | **92.4** | 45.1 | — | 4.8 |
| $m = 3$ | | 0.0 | 0.0 | 0.0 | 0.0 | 8.3 | **91.6** | 44.4 | — | 4.8 |
| $m = 10$ | % Recommendation | 0.0 | 0.0 | 0.0 | 0.2 | 11.2 | **88.6** | 39.3 | — | 4.4 |
| Power prior | | | | | | | | | | |
| $n_c = 1$ | | 0.0 | 0.1 | 0.1 | 0.3 | 9.6 | **89.9** | 43.3 | — | 4.7 |
| $n_c = 3$ | | 0.1 | 0.0 | 0.1 | 0.4 | 10.6 | **88.7** | 43.0 | — | 4.7 |
| $n_c = 10$ | | 0.0 | 0.2 | 1.7 | 4.2 | 12.0 | **81.9** | 38.2 | — | 4.3 |

*(continued)*

**Table 2.** (continued)

| Method | | Dose Level $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | Allocation % MTD | >MTD | Average Toxicity |
|---|---|---|---|---|---|---|---|---|---|---|
| Asian scenario 6 | True toxicity prob. | 0.03 | 0.06 | 0.12 | 0.21 | **0.36** | 0.53 | | | |
| Noninformative prior | | 0.0 | 0.5 | 4.5 | 27.9 | **41.5** | 25.6 | 23.6 | 13.7 | 7.0 |
| Effective sample size prior | | | | | | | | | | |
| $m = 1$ | | 0.0 | 0.0 | 2.0 | 31.3 | **49.1** | 17.7 | 26.6 | 13.6 | 7.4 |
| $m = 3$ | | 0.0 | 0.0 | 1.6 | 32.4 | **50.3** | 15.7 | 27.1 | 11.8 | 7.2 |
| $m = 10$ | % Recommendation | 0.0 | 0.0 | 1.0 | 35.9 | **52.2** | 10.9 | 28.8 | 7.6 | 6.9 |
| Power prior | | | | | | | | | | |
| $n_c = 1$ | | 0.4 | 0.5 | 3.8 | 29.0 | **43.5** | 22.8 | 23.4 | 16.6 | 7.5 |
| $n_c = 3$ | | 0.4 | 0.5 | 3.2 | 31.2 | **43.1** | 21.6 | 23.6 | 16.7 | 7.5 |
| $n_c = 10$ | | 0.2 | 0.4 | 3.9 | 35.5 | **41.5** | 18.6 | 22.6 | 14.8 | 7.3 |

Correct selection percentages are given in boldface. MTD, maximum tolerated dose.

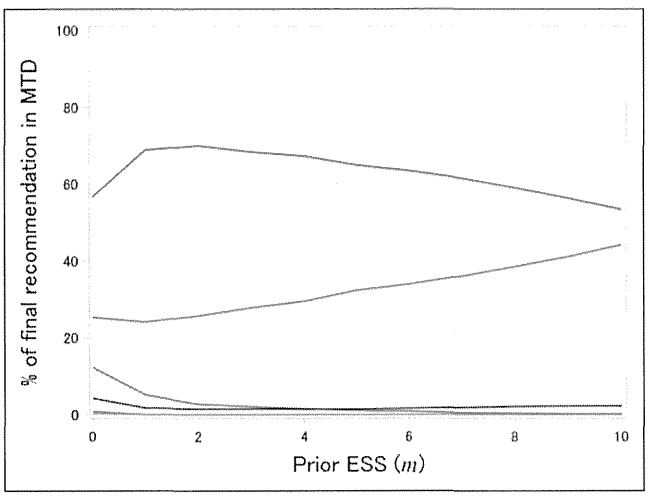


**Figure 1.** Percentages of final recommended MTDs at each dose level ($d_1$: blue, $d_2$: red, $d_3$: green, $d_4$: brown, $d_5$: purple, $d_6$: pale green) with respect to prior ESS values ($m = 0$ to 10) under Caucasian scenario 1 and Asian scenario 3. ESS, effective sample size; MTD, maximum tolerated dose.

previous trials would be useful. It may be possible to improve the robustness of our method by evaluating the interstudy variability of parameters of interest. We could use Bayesian hierarchical models for these purposes.[14] The prior ESS extended to determine the prior informativeness in a conditionally independent hierarchical model[15] may be useful in this setting.

So far, several Bayesian methods have been proposed for evaluating the similarity of treatment effects among patient subgroups in a randomized clinical trial setting.[16,17] Schoenfeld et al[18] proposed a Bayesian approach to pediatric trial design, which allows borrowing strength from previous or simultaneous adult trials. Taking into consideration that pediatric clinicians often rely on evidence from clinical trials in adults, our proposed method can be applied to a dose-finding study for pediatric cancer by regarding adult patients as in the previous trial. In addition, our proposed method can be extended to all phases of a dose-finding study to incorporate historical data—for example, Asian to Caucasian, preclinical to clinical, monotherapy to combination therapy, and previous regimen to current regimen.

## Acknowledgments

We thank Drs Peter Thall and Peter Müller for their helpful comments and useful suggestions. We also thank the associate editor and the referees for their thoughtful and constructive comments and suggestions.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

Satoshi Morita's work was supported in part by a Grant-in-Aid for Scientific Research (C-24500345) from the Ministry of Health, Labour and Welfare of Japan and by the nonprofit organization Epidemiological and by the nonprofit organization Epidemiological and Clinical Research Information Network.

## References

1. Morita S. Application of the continual reassessment method to a phase I dose-finding trial in Japanese patients: East meets West. *Stat Med.* 2011;30:2090-2097.
2. Ogura T, Morita S, Yonemori K, et al. Exploring ethnic differences in toxicity in early-phase clinical trials for oncology drugs [published online March 3, 2014]. *Therapeutic Innovation & Regulatory Science.*
3. Budman DR, Meropol NJ, Reigner B, et al. Preliminary studies of a novel oral fluoropyrimidine carbamate: capecitabine. *J Clin Oncol.* 1998;16:1795-1802.
4. Mackean M, Planting A, Twelves C, et al. Phase I and pharmacologic study of intermittent twice-daily oral therapy with capecitabine in patients with advanced and/or metastatic cancer. *J Clin Oncol.* 1998;16:2977-2985.

5. Pharmaceuticals Medical Devices Agency. Review reports (capecitabine) [in Japanese]. http://www.info.pmda.go.jp/shinyaku/P200700068/45004500_21500AMZ00400_A100_1.pdf. Accessed March 25, 2014.
6. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 1990;46:33-48.
7. Thall P, Lee S-J. Practical model-based dose-finding in phase I clinical trials: method based on toxicity. *Int J Gynecol Cancer*. 2003;13:251-261.
8. Carter SK. Study design principles for the clinical evaluation of new drugs as developed by the chemotherapy program of the National Cancer Institute. In: Staquet MJ, ed. *The Design of Clinical Trials in Cancer Therapy*. Brussels, Belgium: Editions Scientifiques Europrennes; 1973:242-289.
9. Morita S, Thall PF, Mueller P. Determining the effective sample size of a parametric prior. *Biometrics*. 2008;64:595-602.
10. Morita S, Thall PF, Mueller P. Evaluating the impact of prior assumptions in Bayesian biostatistics. *Stat Biosci*. 2010;2:1-17.
11. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15:46-60.
12. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. London, England: Chapman & Hall; 1996.
13. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Stat Med*. 2009;28:3562-3566.
14. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13:41-54.
15. Morita S, Thall PF, Mueller P. Prior effective sample size in conditionally independent hierarchical models. *Bayesian Anal*. 2012; 7:591-614.
16. Liu JP, Hsiao CF, Hsueh H. Bayesian approach to evaluation of bridging studies. *J Biopharm Stat*. 2002;12:401-408.
17. Goodman SN, Sladky JT. A Bayesian approaches to randomized controlled trials in children utilizing information from adults: the case of Guillain-Barre syndrome. *Clin Trials*. 2005;2:305-310.
18. Schoenfeld DA. Bayesian design using adult data to augment pediatric trials. *Clin Trials*. 2009;6:297-304.

Research Article

Statistics in Medicine

Received 16 August 2013, Accepted 23 April 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6209

# Biomarker-based Bayesian randomized phase II clinical trial design to identify a sensitive patient subpopulation

**Satoshi Morita,[a*†] Hideharu Yamamoto[b] and Yasuo Sugitani[b]**

**The benefits and challenges of incorporating biomarkers into the development of anticancer agents have been increasingly discussed. In many cases, a sensitive subpopulation of patients is determined based on preclinical data and/or by retrospectively analyzing clinical trial data. Prospective exploration of sensitive subpopulations of patients may enable us to efficiently develop definitively effective treatments, resulting in accelerated drug development and a reduction in development costs. We consider the development of a new molecular-targeted treatment in cancer patients. Given preliminary but promising efficacy data observed in a phase I study, it may be worth designing a phase II clinical trial that aims to identify a sensitive subpopulation. In order to achieve this goal, we propose a Bayesian randomized phase II clinical trial design incorporating a biomarker that is measured on a graded scale. We compare two Bayesian methods, one based on subgroup analysis and the other on a regression model, to analyze a time-to-event endpoint such as progression-free survival (PFS) time. The two methods basically estimate Bayesian posterior probabilities of PFS hazard ratios in biomarker subgroups. Extensive simulation studies evaluate these methods' operating characteristics, including the correct identification probabilities of the desired subpopulation under a wide range of clinical scenarios. We also examine the impact of subgroup population proportions on the methods' operating characteristics. Although both methods' performance depends on the distribution of treatment effect and the population proportions across patient subgroups, the regression-based method shows more favorable operating characteristics. Copyright © 2014 John Wiley & Sons, Ltd.**

**Keywords:** biomarker; molecular-targeted agent; Bayesian statistics; randomized phase II trial; time-to-event data

## 1. Introduction

Recently, the benefits and challenges of incorporating biomarkers into the development of anticancer agents have been increasingly discussed [1]. Many clinical trials are conducted to develop new molecular-targeted anticancer agents that are likely to benefit only a subset of patients. If a clinical trial is performed in a broad population of patients, which includes insensitive as well as sensitive patients, any effect of a new agent on the sensitive subset of patients may be missed. Therefore, drug development should aim to optimize the target population of patients for treatment by appropriately focusing on patients who could obtain a sufficient benefit from a molecular-targeted agent. In addition, identifying the sensitive subset of patients may be a vital process in clinical development in terms of speeding up the drug development process and reducing development costs [2–5].

The following two examples of clinical development represent two different extremes in the approach to this problem. First, trastuzumab, which is a humanized monoclonal antibody with high specificity for the human epidermal growth factor receptor 2 (HER2) protein, demonstrated high antitumor activity in patients with HER2-overexpressing metastatic breast cancer [6–8]. Based on preclinical and clinical data that strongly supported the existence of a sensitive subpopulation of patients, the clinical development of

[a] *Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan*
[b] *Clinical Research Planning Department, Chugai Pharmaceutical Co., Ltd., Tokyo, Japan*
**Correspondence to: Satoshi Morita, Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan.*
*† E-mail: smorita@kuhp.kyoto-u.ac.jp*

trastuzumab prospectively focused on studying the agent in HER2-overexpressing breast cancer patients. Secondly, during the development of monoclonal antibodies targeting epidermal growth factor receptor (EGFR), such as panitumumab, and EGFR tyrosine kinase inhibitors, such as gefitinib, patients were enrolled in clinical trials without preselection based on EGFR status or other biomarkers [6, 7]. For example, Amado *et al.* [9] retrospectively analyzed whether the effect of panitumumab on progression-free survival (PFS) in patients with metastatic colorectal cancer differed by KRAS status and showed a significant treatment effect in the wild-type KRAS subgroup. That is, in the first case, solid prior data enabled clinical investigators to prospectively design subsequent clinical trials to develop a molecular-targeted agent in a patient subpopulation identifiable with a biomarker assay. In the other case, retrospective subgroup analysis of a phase III trial conducted in unselected patients was able to successfully identify a sensitive patient subpopulation. In many cases, however, the reality may lie in between these two cases.

If a study population of patients contains nonsensitive subpopulations, a much larger sample size would be required to establish statistically significant results in a final confirmatory phase III trial [10]. When considering the entire course of a new agent's clinical development, therefore, conducting a properly designed phase II trial may be key to raising the 'success probability' of a subsequent phase III trial. In particular, pharmacogenetically developed drugs often rely on assays to measure target expression levels (e.g., HER2 or EGFR) on a graded scale; these levels are then dichotomized to define two subsets of patients with positive or negative status. We call the subset of patients with positive status the sensitive subpopulation. In this paper, we consider identifying the sensitive subpopulation using a graded-scale biomarker in a randomized phase II clinical trial to develop a new molecular-targeted agent. In order to design the phase II trial, we adopt a Bayesian approach for the decision-making flexibility it affords during the exploratory phase of clinical development. We compare two Bayesian methods, one based on subgroup analysis and the other on a regression model, in terms of their performance in identifying a sensitive subpopulation. In addition, we consider interim analyses to prematurely terminate the trial because of futility.

As reviewed by Yin [11], there is a substantial literature on study designs that are used to identify sensitive patient subpopulations, including Jiang *et al.* [10], Wang *et al.* [12], Brannath *et al.* [13], and Eickhoff *et al.* [14], and Jenkins *et al.* [15] proposed adaptive two-stage designs in which the patient subset(s) specified in the first stage is used to evaluate the treatment effect in the second stage. Their proposed study designs presume that two mutually exclusive patient subgroups are determined in advance on the basis of preclinical research or a separate exploratory study. Our focus is simply on identifying a sensitive patient subpopulation in the phase II stage, although the preceding study designs consider phase II/III or phase III trial settings.

This paper is organized as follows. In Section 2, we provide a motivating example. Section 3 outlines the study design of a Bayesian randomized phase II clinical trial to identify a sensitive patient subpopulation. We conduct extensive simulation studies to examine the operating characteristics of our proposed study design in Section 4. We close with a brief discussion in Section 5.

## 2. A motivating example

In this section, we present a case study based on the actual clinical development of a new molecular-targeted monoclonal antibody. Preclinical and clinical works suggested that antitumor activity of the new antibody should depend significantly on the target protein amounts. In this study, the intensity of the biomarker expression is defined using a graded scale (e.g., 0, 1+, 2+, and 3+), with higher values indicating higher expression. Results from a phase I dose-finding clinical trial suggested a possible association between biomarker expression and the efficacy of the antibody, that is, a longer PFS time tended to be observed in patients with a higher expression (e.g., 2+ and 3+). In this study, we assume monotonicity in the efficacy of the new agent with respect to the biomarker grade.

While effective first-line therapies exist for patients with advanced stages of cancer and poor prognoses, in particular hepatocellular carcinoma (HCC) and pancreatic carcinoma, no standard second-line treatments have yet been established. In randomized phase II clinical trials to develop second-line oncology treatments, the experimental and control arms (arms E and C) should be the 'best supportive case (BSC) + new agent' and 'BSC + placebo', and a time-to-event outcome such as PFS time is often used as the primary endpoint [16]. In some cases, a biomarker may not only be a predictive factor for a new agent but also a prognostic factor for patients with a specific cancer type. In this study, we assume that the biomarker predicts the efficacy of the new agent but does not predict patient prognosis. That is, we

consider the situation where the efficacy in the control (placebo) arm is not modified by the biomarker. However, it is not difficult to extend our proposed study design to cases where prognosis differs between subgroups.

Under these settings, we consider designing a randomized phase II trial to assess whether the addition of a new monoclonal antibody therapy to BSC sufficiently benefits the patients in terms of prolongation of PFS time. The biomarker grade is used as a stratification factor when randomization is carried out. In order to summarize the PFS data, we basically use a hazard ratio comparing arm E with arm C, which is denoted by $\lambda$. In this study, we consider evaluating the hazard ratios in $G$ biomarker subgroups, which are denoted by $\lambda_g$, $g = 1, \ldots, G$. Our specific goal is to find the upper subset consisting of subgroups $g \geqslant \kappa_0$, which meets the definition of the sensitive subpopulation, by evaluating these hazard ratios. Then, a subsequent phase III trial is to be conducted in the identified subpopulation. The value of cutoff $\kappa_0 \in \{1, \ldots, G+1\}$ is unknown and will be determined based on data observed in the trial. As one of the two extreme cases, $\kappa_0 = 1$ suggests that arm E should be beneficial for the entire population of patients, and one can make a decision to proceed to a subsequent phase III trial that enrolls the entire population of patients. On the other hand, the cutoff $\kappa_0 = G + 1$ indicates that arm E will not be beneficial for any subgroup and that the 'no-go' decision to a subsequent phase III trial should be taken.

## 3. Biomarker-based Bayesian randomized phase II study design

We use the two Bayesian methods that are both based on a common probability model for PFS time. One method is based on a subgroup analysis (S-A method), and the other on a regression model (R-M method).

### *3.1. Notation, probability model for progression-free survival time, and Bayesian posterior computation*

For patient $i$, let $x_i$ denote the treatment indicator, with $x_i = 1$ if patient $i$ receives the experimental arm and $x_i = 0$ if he or she receives the control arm. Let $T_i$ denote PFS time for patient $i$. For subgroups 1 to $G$ defined by the biomarker grade, $z_{i,g} = 1$ if patient $i$ is in subgroup $g$ and 0 if not. Thus, $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,G})$ is the subgroup indicator vector for patient $i$. Let $\phi_1, \ldots, \phi_G$ denote the proportions of patients in subgroups $1, \ldots, G$, who would be enrolled into the phase II trial. These proportions reflect the true biomarker subgroup prevalence in the entire population of patients. Although $\Phi = (\phi_1, \ldots, \phi_G)$ is actually unknown, in the simulation study, we will handle the proportions $\Phi$ as fixed values and vary the values to examine the sensitivity of simulation results to the subgroup prevalence. That is, although the proportions $\Phi$ could be handled as additional parameters to be estimated in a Bayesian study design, we will not consider them in this study.

The two Bayesian methods explained in the next subsection commonly use the following proportional hazards model. Under the proportional hazards assumption in each subgroup, the hazard at time $t$ for patient $i$ with $x_i$ can be modeled as

$$h\left(t \mid x_i, \mathbf{z}_i\right) = h_0(t) \exp\left(\sum_{g=1}^{G} \beta_g x_i z_{i,g}\right), \tag{1}$$

where $h_0(t)$ denotes the baseline hazard function and $\beta_g$ denotes the regression coefficient for $x_i$ in subgroup $g$. According to Sinha *et al.* [17] and Ibrahim *et al.* [18], we use the partial likelihood of the Cox proportional hazards model as the likelihood to compute the posterior distributions of the parameters in the two Bayesian methods. We used Markov chain Monte Carlo to compute the posteriors [19], because the joint posterior distribution of regression coefficient parameters is not readily available in closed form.

As the criteria to identify the sensitive subpopulation, we basically use the following Bayesian posterior probability given the observed data $\mathcal{D}$ from the trial,

$$p\left(\lambda < \eta^* \mid \mathcal{D}\right) > \pi^*, \tag{2}$$

where $\eta^*$ is the upper limit and $\pi^*$ is the upper probability cutoff. These design parameters, $\eta^*$ and $\pi^*$, need to be calibrated on the basis of operating characteristics of the study design, which are examined in

simulation studies. More specifically, let $\mathcal{D}_g$ denote the data observed in subgroup $g$ and $\mathcal{D}_{all}$ denote the data observed in all $G$ subgroups.

### 3.2. *Two Bayesian methods to analyze progression-free survival time*

The objective of the phase II trial is to prove the concept of a targeted therapy, that is, to evaluate whether higher efficacy of the new antibody is observed in patients with higher biomarker expression. Therefore, we assume the monotonicity in the efficacy of the new antibody in both methods but in different ways.

The S-A method separately evaluates the hazard ratio in each subgroup using the data observed in that subgroup. Assuming the monotonic increase in $p(\lambda_g < \eta^* \mid \mathcal{D}_g)$ for $g = 1, \ldots, G$, this method sequentially assesses whether $p(\lambda_g < \eta^* \mid \mathcal{D}_g) > \pi^*$ from subgroups 1 to $G$. That is, if $p(\lambda_1 < \eta^* \mid \mathcal{D}_1)$ is higher than $\pi^*$, we determine $\kappa_0 = 1$. If not, we proceed to subgroup 2. If $p(\lambda_2 < \eta^* \mid \mathcal{D}_2) > \pi^*$, we determine $\kappa_0 = 2$ and decide to identify subgroups 2 to $G$ as the sensitive subpopulations. Similar computations and decision making are then repeated up to subgroup $G$. If all of the posterior probabilities, $p(\lambda_1 < \eta^* \mid \mathcal{D}_1)$, $\ldots$, $p(\lambda_G < \eta^* \mid \mathcal{D}_G)$ are lower than $\pi^*$, we determine $\kappa_0 = G + 1$. We assume a noninformative normal prior N(0, 1000) for each of the regression coefficient parameters, $\beta_1, \ldots, \beta_G$, to perform these posterior computations.

The R-M method assumes a monotonic decrease in hazard ratio for the biomarker subgroups with the parameter constraint $\beta_1 > \beta_2 > \cdots > \beta_G$. In addition, this method uses the data observed in all $G$ subgroups, $\mathcal{D}_{all}$, to evaluate the posterior distribution of $\lambda_g$ for $g = 1, \ldots, G$. For computational convenience, we reparameterize $(\beta_1, \ldots, \beta_G)$ with $(\beta_1, \gamma_1, \ldots, \gamma_{G-1})$ as $\beta_1 = \beta_1$, $\beta_2 = \beta_1 - \gamma_1$, $\ldots$, $\beta_G = \beta_{G-1} - \gamma_{G-1} = \beta_1 - \gamma_1 - \gamma_2 - \cdots - \gamma_{G-1}$, where $\gamma_1 > 0, \gamma_2 > 0, \ldots, \gamma_{G-1} > 0$. Assuming a noninformative normal prior N(0, 1000) for $\beta_1$ and a noninformative gamma prior Ga(0.001, 0.001) with mean 1 and variance 1000 for $\gamma_1, \ldots, \gamma_{G-1}$, we compute the marginal posterior distribution of the hazard ratios. Based on the computations, we find the cutoff $\kappa_0$ to satisfy the following equation:

$$\kappa_0 = \inf_{g \in (1,\ldots,G)} \left\{ g \mid p\left(\lambda_g < \eta^* \mid \mathcal{D}_{all}\right) > \pi^* \right\}. \tag{3}$$

That is, the cutoff $\kappa_0$ is specified as the minimum of the integers $g \in \{1, \ldots, G\}$ that meet $p(\lambda_g < \eta^* \mid \mathcal{D}_{all}) > \pi^*$.

Although we suppose the S-A method has more flexibility, it may perform more poorly at identifying a sensitive subpopulation because of its S-A approach. In contrast, although we expect the R-M method to show a higher performance owing to the parameter constraint and the use of $\mathcal{D}_{all}$, this method may be vulnerable to departures from the monotonicity assumption. We will evaluate the advantages and disadvantages of the two methods in the simulation study.

### 3.3. *Interim study monitoring rules*

It may be important to terminate a clinical trial early from ethical and practical points of view. In the randomized phase II trial, we consider early termination of the entire trial due to futility by planning interim analyses.

Although it may also be useful to consider partly terminating insensitive patient subgroups or reducing the size of those subgroups, we did not take these measures in this study. This is because it may be generally desirable to obtain sufficient data on patients in the nonselected subpopulation in order to more precisely evaluate their response to and the safety of the new treatment [20].

The number and timing of interim analyses should be determined by taking into account the practicalities of patient enrollment rates and collecting and processing of study data. In the randomized phase II trial, we consider two interim analyses with the first and second analyses occurring after 60% and 80% of patients are recruited, respectively. When using the S-A method, given the lower probability cutoff $\pi^*_{stop}$, we consider the experimental arm to have disappointingly insufficient efficacy if $p(\lambda_g < \eta^* \mid \mathcal{D}_g) < \pi^*_{stop}$ for all $g$. Similarly, we stop the trial early if $p(\lambda_g < \eta^* \mid \mathcal{D}_{all}) < \pi^*_{stop}$ for all $g$ when using the R-M method. The lower cutoff $\pi^*_{stop}$ needs to be calibrated on the basis of the study design operating characteristics in the same way as the upper cutoff $\pi^*$. As another interim monitoring rule, it may be useful to include early stopping for efficacy by using an efficacy stopping criterion, such as $p(\lambda_g < \eta^* \mid \mathcal{D}) > \pi^*_{stop,Eff}$. Owing to the same reasons mentioned earlier, however, we will not apply this rule to the phase II trial.

## 4. Evaluation of operating characteristics

### *4.1. Parameter calibration and simulation plan*

To evaluate and compare the two Bayesian methods in the case study with four subgroups, we simulated the trial 5000 times using extensively varying situations. We used Markov chain Monte Carlo methods to obtain samples from the posterior distributions of the parameters. In order to complete the study design, we needed to calibrate the design parameters $(\eta^*, \pi^*, \pi^*_{stop}, N)$ on the basis of the desired type I error rate under a null hypothesis and power under an alternative hypothesis in the trial with the projected total sample size $N$. The detailed definitions of type I error and power are given in the following.

We first performed a series of simulation studies with all 12 combinations of the three fixed upper limits ($\eta^* = 0.70, 0.80, 0.85$), the two upper probability cutoffs ($\pi^* = 0.70, 0.80$), and the two lower probability cutoffs ($\pi^*_{stop} = 0.10, 0.20$) under $N = 500$. Although the total sample size of 500 may be too large for a phase II trial, we used $N = 500$ to reliably evaluate the performances of the two methods in the simulation study. The simulation results are summarized in supplemental tables (see the supporting information). After determining the best combination of $\eta^*$, $\pi^*$, and $\pi^*_{stop}$, we evaluated the operating characteristics using six sample size values ($N = 250, 300, 350, 400, 450$, and 500) to determine the appropriate sample size for the randomized phase II trial. Furthermore, we assumed the five patterns of subpopulation proportions $\Phi = (\phi_1, \phi_2, \phi_3, \phi_4)$, as shown in Table I, to evaluate the sensitivity of simulation results to the subgroup prevalence. We predicted that patterns 1 and 3 were more likely to be observed in the phase II trial according to the historical data.

We assumed the five clinical scenarios for the simulation study based on hazard ratios as shown in Table I. Each scenario is characterized by the true (fixed) hazard ratios ($HR_1$, $HR_2$, $HR_3$, $HR_4$) for the four subgroups. Scenario (1) is a null case, with all hazard ratios equal to 1.0. The sensitive subpopulation, found under each scenario, is indicated in boldface. In order to define the sensitive subpopulation, we first specify the efficacy threshold so that subgroup $g$ is contained in the sensitive subpopulation if $HR_g \leqslant$ the threshold. One possible way to specify the efficacy threshold may be to hold discussions with physicians regarding the published results of clinical trials, because such a specification needs to take into account the current medical environment, such as state-of-the-art therapy and medical costs. For example, in advanced HCC, Llovet *et al.* [21] explored the ability of several biomarkers to predict the efficacy of a new small molecule, sorafenib, using the data from the phase III sorafenib HCC assessment randomized protocol trial [22]. Based on this report as well as other previous data, we solicited the opinions of the two hepatologists in the study group regarding the efficacy threshold. They suggested that an efficacy threshold of 0.6 should be clinically acceptable. We will use a power value to designate the probability of correctly identifying the target subgroup(s) as the sensitive subpopulation under alternative scenarios and a type I error to designate the probability of identifying any subgroup(s) under the null scenario.

**Table I.** Patient subgroup population proportions $\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$ and clinical scenarios characterized by the true (fixed) hazard ratios $\{HR_1, HR_2, HR_3, HR_4\}$ for subgroups 1–4 for the simulation study.

| | | Subgroup | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Subpopulation proportion patterns | | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\phi_4$ |
| 1 | Equal | 0.25 | 0.25 | 0.25 | 0.25 |
| 2 | Higher in subgroups 1 and 4 | 0.35 | 0.15 | 0.15 | 0.35 |
| 3 | Higher in subgroups 2 and 3 | 0.15 | 0.35 | 0.35 | 0.15 |
| 4 | Increasing | 0.05 | 0.15 | 0.30 | 0.50 |
| 5 | Decreasing | 0.50 | 0.30 | 0.15 | 0.05 |
| Clinical scenarios | | $HR_1$ | $HR_2$ | $HR_3$ | $HR_4$ |
| (1) | Null case | 1.0 | 1.0 | 1.0 | 1.0 |
| (2) | Linear | 1.0 | 0.8 | **0.6** | **0.4** |
| (3) | Step-down | 1.0 | **0.6** | **0.6** | **0.35** |
| (4) | High efficacy in subgroups 3 and 4 | 1.0 | 1.0 | **0.5** | **0.3** |
| (5) | High efficacy only in subgroup 4 | 1.0 | 1.0 | 1.0 | **0.5** |

The hazard ratio values in the sensitive subpopulation under each scenario are indicated in boldface.

Taking historical data on second-line therapies for HCC into account, for the simulations, we assumed that the median PFS time was 2.8 months for all four subgroups in the control arm of the trial, with 12.0 months of patient recruitment and 15.0 months of maximum follow-up (i.e., 3.0 months of minimum follow-up). In addition, we assumed that patients arrived uniformly during the recruitment period. Assuming that the patient PFS times are independent and identically distributed exp($\nu$), exponential with parameter $\nu$, which has a PDF of $f(t \mid \nu) = \nu \exp(-\nu t)$, we generated PFS times using the fixed parameter $\nu_c = 0.33$ for the control arm. For the experimental arm, we used the parameter $\nu_c HR_g$ to generate PFS times in subgroups $g$ for $g = 1, \ldots, 4$. The SAS programs to carry out simulations using the S-A and R-M methods are provided in the supporting information (SAS for Windows release 9.3, SAS Institute Inc., Cary, NC, USA).

### 4.2. *Simulation results*

In presenting the results of the simulation studies comparing the S-A and R-M methods, we summarize the probabilities of identifying the following: (i) none of the four subgroups; (ii) subgroup 4 only; (iii) subgroups 3 and 4; (iv) subgroups 2–4; and (v) all four subgroups, as being in the sensitive subpopulation; these categories are denoted by $\mathcal{P}_{none}$, $\mathcal{P}_4$, $\mathcal{P}_{3-4}$, $\mathcal{P}_{2-4}$, and $\mathcal{P}_{all}$, respectively. We chose the combination of $\eta^* = 0.80$, $\pi^* = 0.70$, and $\pi^*_{stop} = 0.2$, which were judged to provide the best operating characteristics for the two methods, based on the extensive simulations (as shown in the supplementary tables in the supporting information). Table II shows the simulation results with $N = 500$ under the five clinical scenarios with the five patterns of patient subpopulation proportions.

Under scenario 1 (null), the R-M method yielded extremely high probabilities of identifying none of the four groups ($\mathcal{P}_{none} = 0.98$–$1.00$), while the values of $\mathcal{P}_{none}$ with the S-A method were 0.70–0.80. That is, the R-M method sufficiently controlled type I error, holding it to less than 0.05 regardless of the pattern of subpopulation proportions under $N = 500$, while the S-A method did not. In addition, the R-M method resulted in early trial termination due to considerably high probabilities of identifying none of the four groups, especially at the first interim analysis. The likelihood of early termination differed significantly between the R-M and S-A methods. This may be because the R-A method analyzed the data observed in all four subgroups, resulting in much sharper posterior distributions of $\lambda_g$ than those obtained by the S-A method, which used the data observed in each subgroup.

Under scenario 2 (linear), neither of the two methods worked sufficiently well; that is, $\mathcal{P}_{3-4}$ were at most 0.50 for both methods. In cases where an obvious sensitive subpopulation may not seem to exist, such as in a scenario that assumes that the hazard ratios change steadily over subgroups, it may be hard to definitively identify the target subpopulation using either of the methods. Under scenario 3 (step-down), although both the S-A and R-M methods performed well overall, the performance of the R-M method may depend significantly on subpopulation proportions. In pattern 4 in particular, where the number of patients enrolled in subgroup 1 (nonsensitive subpopulation) was very slight, the R-M method was more likely to select all the subgroups, resulting in poorer performance. Under scenario 4 (very high efficacy in subgroups 3 and 4), the R-M method selected subgroups 3 and 4 at sufficiently high probabilities across all patterns of subpopulation proportions, and these probabilities were higher than or almost equal to those obtained by the S-A method. Under scenario 5 (very high efficacy only in subgroup 4), the two methods were almost comparable in terms of the probability of identifying subgroup 4 under pattern 1. In cases where the subpopulation proportion of subgroup 4 (sensitive subpopulation) was relatively high, such as in patterns 2 and 4, the R-M method performed much better than the S-A method, as expected. However, under patterns 3 and 5, in which the subpopulation proportion of subgroup 4 was small, the performance of the R-M method was lower than that of the S-A method.

Figure 1 indicates the type I error rates (lower circles) and power values (upper circles) provided by the R-M method for the six sample sizes ($N = 250, 300, 350, 400, 450,$ and 500) under the five patterns of subpopulation proportions. In this simulation study, we focused only on the R-M method because the S-A method could not sufficiently control the type I error rate even under $N = 500$. The R-M method held the type I error to less than 0.05 even under $N = 250$. In terms of providing 80% of the power, $N = 300$ may be sufficient for the projected total sample size of the phase II trial, considering that we actually expect the subpopulation proportions to be like pattern 1 or 3.