University of Colorado at Denver and Health Sciences Center

Effect of Length Biased Sampling of Unobserved Sojourn Times on the Survival Distribution When Disease is Screen-Detected

Karen Kafadar, Philip C. Prorok

May 2007

UCDHSC/CCM Report No. 244

CENTER FOR COMPUTATIONAL MATHEMATICS REPORTS

Effect of Length Biased Sampling of Unobserved Sojourn Times on the Survival Distribution When Disease is Screen-Detected

Karen Kafadar
Department of Mathematics
University of Colorado–Denver
and Health Sciences Center
Denver, Colorado 80217-3364

Philip C. Prorok
Biometry Research Group
Division of Cancer Prevention
National Cancer Institute
Bethesda, Maryland 20892-7354

Abstract

Data can arise as a length-biased sample rather than as a random sample; e.g., a sample of patients in hospitals or of network cable lines (experimental units with longer stays or longer lines have greater likelihoods of being sampled). The distribution arising from a single length-biased sampling time has been derived (e.g., Cox and Lewis 1972) and applies when the observed outcome relates to the random variable subjected to length-biased sampling. Zelen (1976) noted that cases of disease detected from a screening program likewise form a length-biased sample among all cases, since longer sojourn times afford greater likelihoods of being screen-detected. In contrast to the samples on hospital stays and cable lines, however, the length-biased sojourns (preclinical durations) cannot be observed, although their subsequent clinical durations are. This paper quantifies the effect of length-biased sampling of the sojourn times (or pre-clinical durations) on the distribution of the observed clinical durations when cases undergo periodic screening for disease. We show that, when preclinical and clinical durations are positively correlated, the mean clinical duration can be substantially inflated — even in the absence of any benefit on survival from the screening procedure.

Screening studies that report mean survival time need to take account of the fact that, even in the absence of any real benefit, the mean survival among cases in the screen-detected group will be longer than that among interval cases or among cases that arise in the control arm, above and beyond lead time bias, simply by virtue of the length-biased sampling phenomenon.

Key words: randomized screening trial, sojourn time, lead time, bivariate gamma distribution, clinical duration, periodic screening, HIP trial

1 Introduction

Well-designed randomized screening trials provide useful evaluations of the effectiveness of screening programs for the early detection of a chronic disease such as cancer. In such trials, participants in the study arm are offered periodic screening (e.g., five annual screens) and participants in the control arm are told to follow their "usual medical care." At the end of the trial, one may compare either the mortality rates from the disease, or the survival times of cases of disease detected by screening with those from non-screen-detected cases. The breast cancer screening trial conducted by the Health Insurance Plan (HIP) of New York evaluated the effectiveness of mammography plus clinical breast exam in the 1960s; Shapiro et al. (1988) estimated the reduction in breast cancer mortality, averaged among all women in the trial, as roughly 30%. When evaluating screening in terms of the survival times, typically the survival times are measured from the time of start of study as the origin, rather than from the time of diagnosis, because lead time (time by which the diagnosis is advanced, even in the absence of any real benefit in terms of increased survival time) does not affect the comparison (Morrison 1979, Shapiro et al. 1988). However, length biased sampling affects the evaluation, whether the effectiveness is measured either by percent reduction in mortality or extended survival time (Zelen and Feinlieb 1969, Zelen 1976): cases with longer sojourn times are more likely to be detected by screening than those with shorter sojourns. (Another term

for "sojourn time" is "preclinical duration;" both terms are used interchangably in this paper.) Unlike other potential biases potentially affecting the evaluation of screening (e.g., overdiagnosis bias, lead time bias), the effect on the survival times from length biased sampling on the preclinical durations cannot be removed by trial design.

We rely on a familiar three-state disease progression model: (1) disease-free state; (2) preclincial state (disease has not surfaced clinically but can be detected by a screening procedure); (3) clinical state (disease confirmed by clinical diagnostic test). Figure 1 shows these three states both for (a) a case that is not subjected to screening and (b) a comparable screen-detected case (b). Conceptually, both cases experience the same disease history (preclinical and clinical durations), but the detection of the screened case is advanced by an amount known as "lead time", resulting possibly in an extension to the clinical duration by an amount known as "benefit time" (Kafadar and Prorok 2001). The durations illustrated in Figure 1 are conceptual, because, as we show below, the average preclinical duration among the screen-detected cases will exceed that among the unscreened cases, due to length biased sampling. Consequently, if sojourn times are positively correlated with clinical durations, screen-detected cases are more likely to have longer clinical durations as well. This article quantifies this increase. The increase depends on the underlying joint density of the preclinical and clinical durations (in the absence of screening), which, unfortunately, cannot be estimated due to unobservable preclinical durations. The joint density can be modeled using a flexible bivariate gamma density function with a wide variety of parameters, from which we can assess both the range of the length-biased effect and the most influential parameters on this effect.

The general case of size-biased sampling arises when the measurement process favors those experimental units whose observations are larger. This phenomenon arises, for example, in surveys of hospital patients (patients whose visits are longer have a greater likelihood of being sampled than those with shorter visits; Wang 1998), product reliability studies using a warranty data base (units

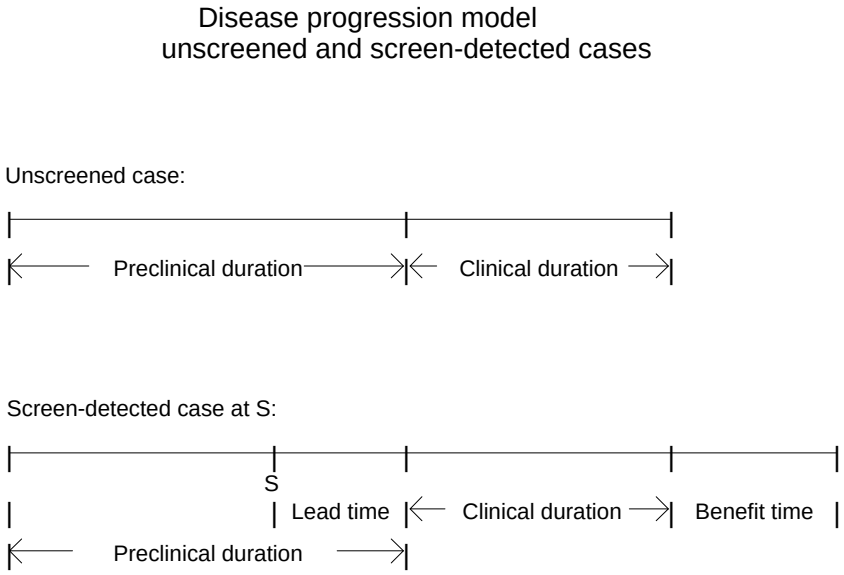

Figure 1: Disease progression model for unscreened and screen-detected cases. Disease-free state precedes preclinical state (not shown). (a) Unscreened case. (b) Comparable screened case: "S" denotes time of screen; L = lead time (between S and point of clinical detection); B = benefit time (extended survival).

in the database or in service for a longer time are more likely to be part of the study; Kalbfleisch, Lawless, and Robinson 1991), or particle densities in a compound (larger particles are more likely to be presented for measurement; Schotz and Zelen 1971). In these cases, the density of the size-biased sample is proportional to both the original (unsampled) density and to its size. Zelen and Feinlieb (1969) noted that length bias enters in the context of screening trials also: screen-detected cases are more likely to have longer sojourn times.

Most of the literature on length-biased sampled data concentrates on statistical methods for estimating the density function (Jones 1991), survival functions (Cox 1969, Vardi 1985), and proportional hazards models (Wang 1996). Wang (1998) briefly describes these methods, as well as

the effect of length biased sampling in models for disease prevalence, truncated data, and recurrent event processes (e.g., length of stays in hospitals). Gupta and Tripathi (1990) analyze the effect of ignoring length bias on the modeling error. Cnaan (1985) proposes a two-phase survival model where the endpoints of both phases are observable and their lengths are exponentially distributed. Blumenthal (1967) estimates the mean lifetime of units sampled on a particular date; such units are length-biased sampled (units that have been in use for longer periods are more likely to be observed, and hence the observations will tend to have longer lifetimes than those in the general population of units). In a relevant paper, Schotz and Zelen (1971) derive joint probability distributions of the durations of the four phases in the life cycle of a proliferating cell, where cells can be labeled only during their second phases (whose durations cannot be measured) and can be observed only during their final phases. Their results apply to our situation when only one screen is offered; below we generalize to the situation involving multiple screens.

Apart from Schotz and Zelen (1971), most of the literature that discusses length-biased sampling does not apply to the disease screening context, because the length-biased sampled variable in a screening trial (i.e., the sojourn time) cannot be observed: the start of the preclinical duration is unobservable (except theoretically, if screening were conducted continuously, which is practically infeasible). In fact, the primary effect of interest caused by length-biased sampling is not on the *preclinical* durations, but rather on the *clinical* durations that follow them. These *are* observable: clinical duration is defined as the duration between the point of clinical detection (in the absence of screening) and the endpoint (e.g., time of cure or of death). Most experts agree that these two durations are positively correlated (Fontana et al. 1991, Spratt et al. 1995, 1996). Consequently, in this paper we quantify the extent to which the length-biased sampling of the sojourn times affects the survival distribution among the screen-detected cases when compared with that among those cases that were not screen-detected. Specifically, we derive the joint density of the preclinical

and clinical duration when the former variable is subject to length-biased sampling, and then calculate, for specific choices of the underlying (unsampled) joint density, the mean increase in clinical duration in the absence of any screening benefit (as measured by increased survival). From this density, we compute means, standard deviations, and percentiles of both variables, and compare them with those from the unsampled joint density. For cases not detected by screening, "survival since diagnosis" is simply the clinical duration. For screen-detected cases, average "survival since diagnosis" can be decomposed into three components: (1) average lead time, or the average time between screen-detection and the time at which detection would have occurred in the absence of screening; (2) average clinical duration, composed of the average clinical duration in the absense of screening plus the average increase in this clinical duration that results from length-biased sampling of the preclinical duration (this paper); and (3) average benefit time, or the average difference between the time of the endpoint in the absence of screening and the endpoint resulting from the screen. A characterization of the screening benefit, unconfounded by the effects of the length biased sampling and lead time, is given by the third component, the benefit time, but "survival since diagnosis" involves all three components. For any given individual, none of these three components can be measured directly. The distribution of clinical durations in the absence of screening, apart from treatment effects, can be estimated from control arm subjects in a randomized screening trial. Models, such as the one used in this paper, facilitate the estimation of the distribution of the remaining components. Kafadar and Prorok (1994; 2005) discuss methods for estimating components (1), lead time, and (3), benefit time; this paper uses a model to estimate means and standard deviations of (2), the effect of length biased sampling.

Several authors have addressed the issue of length biased sampling, and some authors have suggested methods of estimating the bias. Vardi (1982), Vardi (1985), and Gill, Vardi, and Wellner (1988) derived nonparametric maximum likelihood estimates of the underlying, non-length-biased

sampled distribution function. Their work assumes that measurements on the length biased sample are observed directly. Morrison (1979, p. 716) proposed as an estimate of this bias the difference between the mean survival time since diagnosis among the screen detected cases and the sum of the average lead time (which can be estimated by several methods; see the review and comparison in Kafadar and Prorok 1996), and the average clinical duration (which can be estimated as the average survival time since diagnosis among all control arm participants). This method assumes that the average benefit time is zero; i.e., that the screening procedure offers no gain in survival time. If such a benefit exists, then this estimator confounds the benefit time and the length-sampling bias.

In this article, we derive the joint density of the preclinical and clinical durations when the former is subjected to length biased sampling, based on a general nonparametric disease progression model and either a single screening time (Section 2) or multiple periodic screening times (Section 3). Detailed formulae are presented when the underlying joint density is bivariate gamma (nine parameters) in Section 4, with calculations given in Section 5. Sections 6 closes with discussion, conclusions, and plans for future work.

2 Single sampling time

We first recall the length-biased joint density when sampling occurs only once (e.g., with a prevalent screen in the absence of previous screening examinations) on only one of the variables (during the preclinical phase). Let (Y, Z) be the random vector denoting the preclinical and clinical durations in the general population whose disease is not screen-detected. Let (Y^*, Z^*) denote the corresponding durations for screen-detected case; i.e., when Y is subjected to length-biased sampled. Cox and Lewis (1972, p. 67) show heuristically that the probability density function (pdf) of Y^*, $f_{Y^*}(\cdot)$ is related to that of Y, $f_Y(\cdot)$, via

$$f_{Y*}(\cdot) = \lim_{m\to\infty} y \cdot m_y dy / \sum_{i=1}^{m} Y_i = \lim_{m\to\infty} y \cdot (m_y/m) dy / (\sum_{i=1}^{m} Y_i/m) = y \cdot f_Y(y)/\mu_Y$$

where $\{Y_i, i=1,...,m\}$ are the length-biased sampled intervals, $m_y dy$ is the number of the m intervals having lengths in $(y, y+dy)$, and μ_y is the mean of Y, $\int_0^\infty y f_Y(y) dy$. By extension, the joint density $f_{Y^*,Z^*}(y,z)$, when only Y is subject to length-biased sampling, is related to the joint density $f_{Y,Z}(y,z)$, via

$$f_{Y^*,Z^*}(y,z) = f_{Z^*|Y^*}(z|y) f_{Y^*}(y) = f_{Z|Y}(z|y) \cdot y f_Y(y)/\mu_Y = y \cdot f_{Y,Z}(y,z)/\mu_Y. \tag{1}$$

That the conditional distributions of $Z^*|Y^*$ and $Z|Y$ are the same can be proved formally by recognizing that the distribution of Z (or Z^*) depends only on Y (or Y^*), not on the mechanism that caused the length-biased sampling of Y. Hence, letting I denote an indicator variable that takes the value 1 if Y is length-biased sampled and 0 otherwise, $f_{Y^*}(y) = f_{Y|I=1}(y|1)$ and $f_{Z^*|Y^*}(z|y) = f_{Z|Y,I=1}(z|y,1) = f_{Z|Y}(z|y)$ since Z is independent of I. Schotz and Zelen (1971, p.391, eqn 13) prove the equivalent formula when a four-phase process, (X,Y,Z,W), is subjected to length-biased sampling during the second phase, Y. From (1), the proportional increase in the mean of Y^* over the mean of Y, for a single screen, is

$$E(Y^*)/E(Y) = \int_0^\infty \int_0^\infty y^2 f_{Y^*,Z^*}(y,z) dy dz / \mu_Y = (1 + CV_Y^2) \tag{2}$$

where CV_Y denotes the coefficient of variation (relative standard deviation) of Y, as shown also in Cox and Lewis (1972) and in Schotz and Zelen (1971). To evaluate the benefit of screening, the primary interest is in the effect of the length-biased sampling phenomenon on the clinical duration, Z, which, unlike Y, is observable. The mean of Z^* is:

$$\begin{aligned} E(Z^*) &= \int_0^\infty \int_0^\infty z f_{Y^*,Z^*}(y,z) dy dz = \int_0^\infty \int_0^\infty z f_{Z^*|Y^*}(z|y) f_Y^*(y) dy dz \\ &= \int_0^\infty \int_0^\infty z f_{Z|Y}(z|y) \cdot g(y) f(y) dy dz = \int_0^\infty \int_0^\infty g(y) z f_{Y,Z}(y,z) dy dz \\ &= E(g(Y) \cdot Z) \end{aligned} \tag{3}$$

where, in (3), $g(y) = y/\mu_Y$. For a single screen, $f_{Y^*}(y) = y f_Y(y)/\mu_Y$, so the corresponding increase in $E(Z)$ is

$$E(Z^*)/E(Z) = E(YZ)/(\mu_Y \mu_Z) = (\rho \sigma_Y \sigma_Z + \mu_Y \mu_Z)/(\mu_Y \mu_Z) = (1 + \rho CV_Y CV_Z) \tag{4}$$

where μ_Y, σ_Y^2, CV_Y (respectively, μ_Z, σ_Z^2, CV_Z) denote the mean, variance, and relative standard deviation of Y (respectively Z), and ρ is the correlation between Y and Z. Thus, the proportional increase in the average clinical duration among those individuals who are screen-detected, even when screening results in no benefit (extended survival), depends upon ρ and the relative standard deviations. While CV_Z can be estimated from the cases that arise during the control arm of a randomized screening trial, no such estimates of ρ or CV_Y are available. Nonetheless, equation (4) shows that when the CV of the preclinical and clinical durations are 1 ($CV_Y = CV_Z = 1$) and the correlation between them is only 0.5, screening results in an increase of 50% in the apparent clinical duration among the screen-detected cases, even when the screening procedure offers no benefit in terms of increased survival since diagnosis, above and beyond what has been taken into account by lead time. (If $\rho < 0$, $E(Z^*) < E(Z)$, but most experts agree that the correlation between preclinical and clinical durations is unlikely to be negative.) Thus, even in this simple scenario, the effect of length-biased sampling can be significant. Quite possibly ρ could exceed 0.5, resulting in even greater increases in the observed mean clinical duration.

Equation (3) shows that $f_{Y^*}(y)$ can be expressed as a product of two functions, one of which is the underlying probability density function. The next section will show that, when screening is periodic at regular intervals, $f_{Y^*}(y)$ still can be expressed in the form $g(y)f_Y(y)$; the function $g(y)$ is more complicated than y/μ_Y, but the general expression remains the same. Consequently Equation (3) still holds for the mean of Z^*. The variance of both Y^* and Z^* can be calculated directly as

$$Var(Y^*) = E(Y^2 \cdot g(Y)) - [E(Y \cdot g(Y))]^2 \tag{5}$$

$$Var(Z^*) = E(Z^2 \cdot g(Y)) - [E(Z \cdot g(Y))]^2 \tag{6}$$

which will apply to the periodic screening case also.

3 Periodic screening

Equation (4) provides a simple formula for the increase in the mean clinical duration among cases detected at a single prevalence screen. In this section, we derive the density of the length-biased sampled sojourn time, Y^* for both prevalent and incident cases, when screening occurs at regular intervals of length δ. (The effect of test sensitivity, denoted by β, will be incorporated later.) Let X denote the time at which the preclinical phase begins, and let Y and Z denote the durations of the preclinical and clinical phases as before. We suppose that the initial screen occurs at time 0, and hence prevalent cases are detected at that time with probability β. We calculate the pdf of the sampled sojourn times, $f_{Y^*}(y)$, for

- *prevalent* cases that fail to be detected until the k^{th} screen following the initial (baseline) screen, k=1, 2, ...; and
- *incident* cases that arise in the interval $(i\delta,\ j\delta]$ but fail to be detected until the k^{th} screen following the initial (baseline) screen, $i < j < k$.

This terminology is consistent with that used in Miller et al. (2000). From these derivations, we can calculate the mean and variance of the sojourn time for cases detected at any given screen, $f_Y^*(\cdot)$, and then, using equations (3) and (6), the mean and variance of Z^*.

For a stable process, X will be uniformly distributed over some interval. For prevalent cases, this interval may be $[-L,\ 0]$, where L may be many years before screening is initiated; for an incident case that arises immediately following the initial baseline screen, X may be uniformly distributed over $(0,\ \delta]$. We assume that the screening interval δ is constant, though, in practice, δ

is likely to have a distribution (not everyone will be screened at perfectly constant time intervals). The further assumption of a uniform distribution implies that the case arrival process exhibits no particular preference for any time point during which cases may arise. This assumption may be more reasonable when we are considering intervals between two successive screens, but may be less realistic for prevalent cases (where L may be 20 years or more). In any event, the calculations will be independent of the length of this uniform distribution, so consider for simplicity of presentation the situation of an incident case that arises in [0, δ]. Interval cases correspond to the situation where $S \equiv X + Y < \delta$ for which screening does not detect the case and thus had no chance to be effective, so we need be concerned only with the distribution of Y conditional on $S > \delta$, which occurs when the preclinical duration is long enough to cross the screening point at the time (δ) of the first screen following the prevalent screen. Let $Y^*_{(k)}$ be the random variable that denotes the sojourn time of a case that lasted until the k^{th} screen following the interval in which it arose; e.g., the sojourn time for a case that arose in (0, δ) but lasted until time $k\delta$, or a case that arose in (δ, 2δ) but lasted until time $(k+1)\delta$, etc. The asterisk denotes that this sojourn time is eligible for screen-detection and hence is affected by length-biased sampling. Let $F_{Y^*_{(k)}}(y)$ denote cumulative distribution function (cdf) of $Y^*_{(k)}$, given by the conditional probability $\mathrm{P}\{Y \leq y | X + Y > k \cdot \delta\}$. We first calculate this probability when $k = 1$ for two separate regions of y: $y \leq \delta$ and $y > \delta$; see Figures 3(a) and 3(b) respectively. The denominator of this conditional probability is:

$$\begin{aligned} \mathrm{P}\{X + Y > \delta\} &= 1 - \mathrm{P}\{X + Y \leq \delta\} == 1 - \int_0^\delta \int_0^{\delta - s} \delta^{-1} f_Y(t) dt ds \\ &= 1 - \int_0^\delta \delta^{-1} F_Y(\delta - s) ds \equiv 1 - J(\delta)/\delta \end{aligned} \tag{7}$$

where $F_Y(\cdot)$ is the cdf corresponding to the pdf $f_Y(\cdot)$ of the unsampled Y and

$$J(\delta) \equiv \int_0^\delta F_Y(u) du \; . \tag{8}$$

The numerator of the conditional probability of interest when $y \leq \delta$ is (see Figure 3a):

$$\mathrm{P}\left\{X + Y > \delta \text{ and } Y \leq y\right\} = \int_0^y \int_{\delta - t}^{\delta} \delta^{-1} f_Y(t) ds dt = [yF_Y(y) - J(y)]/\delta \tag{9}$$

which leads to

$$F_{Y^*_{(1)}}(y) = [yF_Y(y) - J(y)]/[\delta - J(\delta)]\ ,\ y \leq \delta\,. \tag{10}$$

When $y > \delta$ (see Figure 3b), this probability has two components: that given by Eqn (10) with δ substituted for y, and a second component (for the added probability beyond δ):

$$\begin{aligned} F_{Y^*_{(1)}}(y) &= [\delta F_Y(\delta) - J(\delta)]/[\delta - J(\delta)] + \int_{\delta}^{y} \int_0^{\delta} \delta^{-1} f_Y(t) ds dt/[1 - J(\delta)/\delta]\,,\ y > \delta \\ &= [\delta F_Y(y) - J(\delta)]/[\delta - J(\delta)]\,,\ y > \delta\,. \end{aligned} \tag{11}$$

Note that (10) and (11) are equal when $y = \delta$. Hence the conditional probability density function of the sojourn times for those cases that arose in the screening interval (0, δ], and thus are eligible for screen detection at time δ, is given by:

$$f_{Y^*_{(1)}}(y) = \begin{cases} yf_Y(y)/[\delta - J(\delta)] & y \leq \delta \\ \delta f_Y(y)/[\delta - J(\delta)] & y > \delta \end{cases} \tag{12}$$

The function $f_{Y^*_{(1)}}(\cdot)$ satisfies the properties for a valid density function: $0 \leq J(\delta) \equiv \int_0^{\delta} F_Y(y)dy \leq \int_0^{\delta} 1 dy \leq \delta$, so $0 \leq \delta - J(\delta) \leq \delta$ and hence $f_{Y^*_{(1)}}(y) \geq 0$ for all $y \geq 0$ and $\int_0^{\infty} f_{Y^*_{(1)}}(y)dy = 1$. Both the mean and the variance of $Y^*_{(1)}$ can now be calculated, using (12).

Analogously, the conditional distribution function for $Y^*_{(k)}$, the sojourn time of a case that arose in the first interval following the prevalent screen (0, δ] and went undetected for the prevalent and $(k-1)$ subsequent screens, and lasted at least until the k^{th} screen, at time $k\delta$. If the preclinical phase begins in (0, δ], then it can last until $k\delta$ only if the duration y is at least $(k-1)\delta$. As before, the distribution function is derived when $y \leq k\delta$ and when $y > k\delta$:

$$F_{Y^*_{(k)}}(y) \equiv \mathrm{P}\left\{Y \leq y | X + Y > k\delta\right\}$$

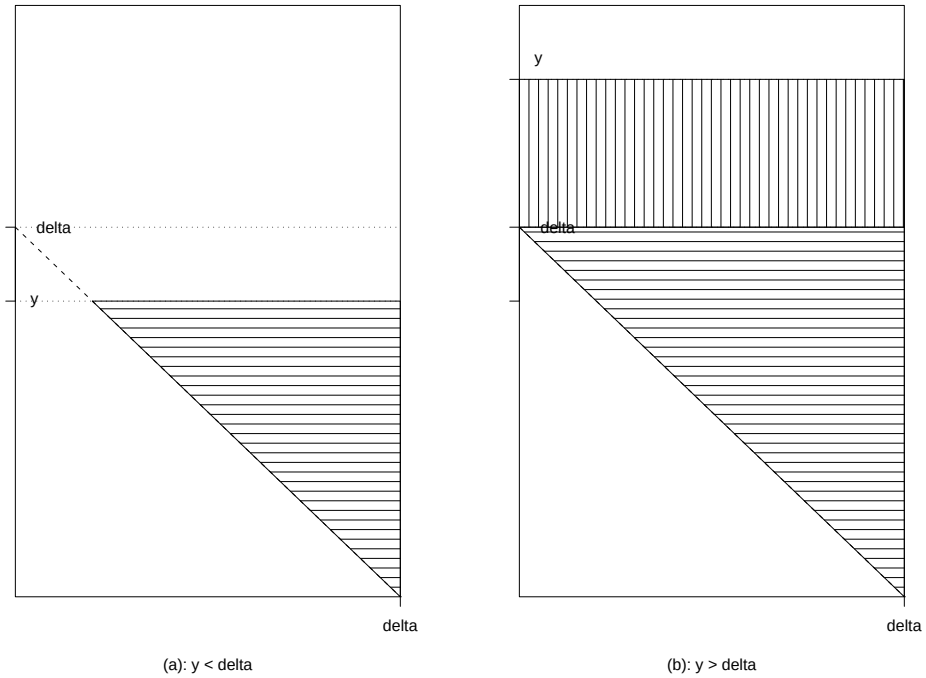

Figure 2: Two regions of integration: (a) region for Eqn (10); (b) region for Eqn (11).

$$= \begin{cases} \{[y-(k-1)\delta]F_Y(y) - [J(y) - J((k-1)\delta)]\} / D_{01k\delta} & (k-1)\delta \le y \le k\delta \\ \{\delta \cdot F_Y(y) - [J(k\delta) - J((k-1)\delta)]\} / D_{01k\delta} & y > k\delta \end{cases} \quad (13)$$

and

$$f_{Y^*_{(k)}}(y) = \begin{cases} [y-(k-1)\delta] \cdot f_Y(y)/D_{01k\delta} & (k-1)\delta \le y \le k\delta \\ \delta \cdot f_Y(y)/D_{01k\delta} & y > k\delta \end{cases} \quad (14)$$

where

$$D_{ijk\delta} \equiv \int_{(k-j)\delta}^{(k-i)\delta} [1 - F_Y(u)]du \equiv (j-i)\delta - [J((k-i)\delta) - J((k-j)\delta)] \quad (15)$$

represents the area under the survival curve of Y between $(k-j)\delta$ and $(k-i)\delta$.

Finally, the same methods provide the cumulative distribution function and probability density function of the sojourn time from an *incident* case that arises in $(i\delta, j\delta]$ and lasts at least until the k^{th} screen following the prevalent screen, at time $k\delta$, $F_{Y^*_{(ij)k}}(\cdot)$:

$$F_{Y^*_{(ij)k}}(y) = \int_{(k-j)\delta}^{y} [F_Y(y) - F_Y(u)]du / D_{ijk\delta} \quad (16)$$

$$= \begin{cases} \{[y-(k-j)\delta]F_Y(y) - [J(y) - J((k-j)\delta)]\} / D_{ijk\delta} & (k-j)\delta < y \leq (k-i)\delta \\ \{[(j-i)\delta]F_Y(y) - [J((k-i)\delta) - J((k-j)\delta)]\} / D_{ijk\delta} & y > (k-i)\delta \end{cases}$$

$$f_{Y^*_{(ij)k}}(y) = \begin{cases} [y-(k-j)\delta]f_Y(y)/D_{ijk\delta} & (k-j)\delta < y \leq (k-i)\delta \\ (j-i)\delta f_Y(y)/D_{ijk\delta} & y > (k-i)\delta \end{cases} \quad (17)$$

These density functions allow the calculation of the pdf of sojourn times (and hence the mean and variance) among cases detected *at* screen k, which fall into three categories:

1. prevalent and screen-detected (with probability β) at the initial screen [the mean of the distribution for these cases was calculated in equation (2) as $\mu_Y(1 + CV_Y^2)$];

2. prevalent at the initial screen, not detected at the initial and $k-1$ subsequent screens (with probability $1-\beta$ at each failed screen), and detected (with probability β) at the k^{th} screen;

3. incident between the i^{th} and j^{th} screens ($i\delta < X \leq j\delta$), not detected at screens $i+1$, ..., $k-1$ (with probability $1-\beta$ at each screen), detected on the k^{th} screen (with probability β; $i < j < k$).

(The third category could apply, for example, to those persons who failed to appear for screening exams $i+1$, ..., $j-1$.) The mean sojourn time among the cases that arise in (0, δ) and are detectable (with probability β) at the time δ of the first screen is:

$$\cdot E(Y^*_{(1)}) = \frac{\int_0^\delta y^2 f_Y(y)dy + \delta \cdot \int_\delta^\infty y f_Y(y)dy}{\delta - J(\delta)} \quad (18)$$

and for those cases whose preclinical durations last until the k^{th} screen (after possibly $k-1$ false negative screens):

$$E(Y^*_{(k)}) = \frac{\int_{(k-1)\delta}^{k\delta} y[y-(k-1)\delta]f_Y(y)dy + \delta \cdot \int_{k\delta}^\infty y f_Y(y)dy}{\delta - J(k\delta) + J((k-1)\delta} \equiv E(Y g_{(k)}(Y)) \quad (19)$$

where the mean is expressed more simply on the right hand side of (19) as

$$g_{(k)}(y) = \frac{[y-(k-1)\delta]I_{((k-1)\delta,k\delta]}(y) + \delta I_{(k\delta,\infty)}(y)}{\delta - [J(k\delta) - J((k-1)\delta]} \quad (20)$$

Here, $I_{(a,b)}(y)$ is the indicator function that takes the value 1 if $y \in (a, b)$ and 0 otherwise. (Note that, for a case that arose in (0,δ) and was screen-detected at screen k, $0 < X < \delta$ and $X + Y > k\delta$ $\Rightarrow Y > (k-1)\delta$.) Equation (14) shows that the density of $Y^*_{(k)}$ can be expressed as a product of y (that depends on k) and the underlying density function of the original unsampled Y: $f^*_{Y_{(k)}}(y) = g_{(k)}(y) f_Y(y)$, where $g_{(k)}(y)$ is given by (20). (Recall that a screen-detected case that arose in (0,δ) and was detected at screen k requires $0 < X < \delta$ and $X + Y > k\delta \Rightarrow Y > (k-1)\delta$.) Consequently, the mean clinical duration among the cases that arose in (0, δ] and are detectable at the k^{th} screen (i.e., whose preclinical duration lasts to the k^{th} screen after $k-1$ false negative screens) is calculated using equation (3) as $E(g_{(k)}(Y)Z)$.

Cases detected *at* the k^{th} screen fall into $k+1$ distinct categories:

(1): cases arose in $((k-1)\delta,\ k\delta]$ and were detected immediately, with probability β, at screen k;

(2): cases arose in $((k-2)\delta,\ (k-1)\delta]$, failed to be detected, with probability $(1-\beta)$, at the first possible screen at time $(k-1)\delta$, but detected, with probability β, at the next screen (at time $k\delta$); ...

(k): cases arose in (0, δ], failed to be detected at the intervening $(k-1)$ screens, with probability $(1-\beta)^{k-1}$, but detected, with probability β, at screen k;

($k+1$): cases arose before time 0, failed to be detected, with probability $(1-\beta)^k$, at either the prevalent screen or any of the $(k-1)$ subsequent screens, but detected, with probability β, at screen k.

The pdf for the sampled sojourn times for cases in categories (1), (2), ..., (k) is:

$$f_{Y^*_{(j)}}(y) = g_{(j)}(y) f_Y(y)\,,\ j = 1, ..., k \tag{21}$$

where

$$g_{(j)}(y) = [(y - (j-1)\delta) \cdot I_{((j-1)\delta, j\delta]}(y) + \delta I_{(j\delta,\infty)}(y)]/D_{01j\delta} \tag{22}$$

and

$$D_{01j\delta} = \delta - [J(j\delta) - J((j-1)\delta)] = \mathrm{P}\left\{(j-1)\delta < X \leq (j-0)\delta,\, X+Y > j\delta\right\}\,. \tag{23}$$

The pdf of the sampled sojourn times for cases in category $(k+1)$ is found analogously from

$$\begin{aligned}
F_{Y^*_{(k^+)}}(y) &= \mathrm{P}\left\{Y \leq y | -L < X < 0\,,\, X+Y > k\delta\right\} &(24)\\
&= \mathrm{P}\left\{Y \leq y\,,\, -L < X < 0\,,\, X+Y > k\delta\right\} \,/\, \mathrm{P}\left\{-L < X < 0\,,\, X+Y > k\delta\right\}\\
&= \int_{k\delta}^{y}\int_{k\delta-v}^{0} f_X(x)f_Y(v)dxdv \,/\, \int_{k\delta}^{\infty}\int_{k\delta-v}^{0} f_X(x)f_Y(v)dxdv\\
&= (1/L)\int_{k\delta}^{y}(v-k\delta)f_Y(v)dv \,/\, (1/L)\int_{k\delta}^{\infty}(v-k\delta)f_Y(v)dv\\
\Rightarrow f_{Y^*_{(k^+)}}(y) &= (y-k\delta)f_Y(y)I_{(k\delta,\infty)}(y) \,/\, D^+_{k\delta} &(25)
\end{aligned}$$

where

$$D^+_{k\delta} \equiv \mu_Y - k\delta + J(k\delta) \;=\; P\{X<0|X+Y>k\delta\}\,. \tag{26}$$

Hence the pdf of sojourn times for cases detected *at* screen k, denoted $f_{Y^*_k}(\cdot)$, is calculated as a weighted average of the conditional densities, where the weights include the probabilities of the conditions and the probabilities of detection:

$$\begin{aligned}
f_{Y^*_k}(y) &= \beta[\sum_{j=1}^{k}(1-\beta)^{j-1}f_{Y^*_{(j)}}(y)D_{01j\delta} \,+\, (1-\beta)^k f_{Y^*_{(k^+)}}(y)D^+_{k\delta}]/w_+\\
&= [\sum_{j=1}^{k} w_j g_{(j)}(y)f_Y(y) \,+\, w_{k^+}g_{(k^+)}(y)f_Y(y)]/w_+\\
&\equiv g_k(y)f_Y(y) &(27)
\end{aligned}$$

where

$$\begin{aligned}
g_k(y) &= [\sum_{j=1}^{k} w_j g_{(j)}(y) \,+\, w_{k^+}g_{(k^+)}(y)]/w_+ &(28)\\
w_j &= \beta(1-\beta)^{j-1}D_{01j\delta}\,,\; j=1,...,k &(29)\\
w_{k^+} &= \beta(1-\beta)^k D^+_{k\delta} &(30)\\
w_+ &= \sum_{j=1}^{k} w_j + w_{k^+} &(31)
\end{aligned}$$

$$g_{(j)}(y) = [(y-(j-1)\delta)I_{((j-1)\delta,j\delta]}(y) + \delta I_{(j\delta,\infty)}(y)]/D_{01j\delta} \quad (32)$$

$$g_{(k^+)}(y) = [(y-k\delta)I_{(k\delta,\infty)}(y)/D^+_{k\delta} \quad (33)$$

Notice that, as $\delta \to 0$, $g_k(y)$ consists of only the final term, $g_{(k+)}(y)$, so $f_{Y^*_k}(y) = y \cdot f_Y(y)/\mu_y$, the length bias sampling effect at the prevalence screen; cf. Eqn (2). Because $D_{01j\delta}$ (respectively, $D_{01k^+\delta}$) in the weights w_j (respectively, w_{k^+}) cancels with the denominators in $g_{(j)}(y)$ (respectively, $g_{(k^+)}(y)$), the pdf $f_{Y^*_k}(y)$ can be written as

$$f_{Y^*}(y) = g_k(y)f_Y(y) \equiv [\sum_{j=1}^{k} \tilde{w}_j \tilde{g}_{(j)}(y) \; + \; \tilde{w}_{k^+}\tilde{g}_{(k^+)}(y)]/w_+ \quad (34)$$

where $\tilde{g}_{(j)}(y) = g_{(j)}(y)D_{01j\delta}$, $\tilde{w}_j = w_j/D_{01j\delta}$, and likewise for $\tilde{g}_{(k^+)}(y)$ and $\tilde{w}_{k^+}$.

From $f_{Y^*}(y)$, Eqn (1) yields $f_{Z^*}(z)$, the pdf of clinical durations for screen-detected cases:

$$\begin{aligned} f_{Y^*,Z^*}(y,z) &= f_{Z^*|Y^*}(z|y)f_{Y^*}(y) \\ &= f_{Z|Y}(z|y)f_{Y^*}(y) \\ &= f_{Z|Y}(z|y)g_k(y)f_Y(y) \\ &= f_{Y,Z}(y,z)g_k(y) \\ \Rightarrow f_{Z^*}(z) &= \int_0^\infty g_k(y)f_{Y,Z}(y,z)\,dy\,. \end{aligned} \quad (35)$$

As above, notice that as $\delta \to 0$, $g_k(y) = y$ and $f_{Z^*_k}(z) = \int_0^\infty y f_{Y,Z}(y,z)dy/\mu_y$; see 4.

4 Specific calculations for gamma-distributed sojourn times

The bivariate gamma density function provides a flexible model for the joint density of Y and Z, the preclinical and clinical durations. By choosing a wide range of possible parameters for this density, we can evaluate the effect of length-biased sampling for a variety of scenarios, using as guidance for these choices the observed behavior of durations for certain diseases (e.g., short preclinical/clincal durations for pancreatic cancer; long durations for prostate cancer). The bivariate gamma density

is a mixture of univariate gamma pdfs, whose cdf has a convenient explicit form, allowing also the explicit calculation of its integral, $J(\cdot)$ (Eqn 8).

The cumulative distribution function (cdf) of the univariate gamma density with parameters r and λ can be computed explicitly and recursively for integral values of r as:

$$\begin{aligned} F_Y(y) &\equiv \Gamma(y;r,\lambda) = \int_0^y \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} dx = \Gamma(\lambda y; r, 1) \\ &= 1 - e^{-\lambda y}\left[1 + \sum_{m=1}^{r-1} (\lambda y)^m / m!\right] \\ &= 1 - e^{-\lambda y}[1 + u_{r-1}(\lambda y)] \end{aligned} \tag{36}$$

where $u_n(x) \equiv \sum_{j=0}^{n} x^j/j! \equiv u_{n-1}(x) + x^n/n!$ (when $n = 1$, $u_1(x) = x$; see also Abramowicz and Stegun 1958, p.262, equations 6.5.13 and 6.5.21). Hence, the integral of the gamma cdf is the sum of integrals of a sequence of gamma cdfs with increasing shape parameters, from which $J(\delta)$ and $D_k \equiv D_{01k\delta}$ can be calculated explicitly via:

$$\begin{aligned} J(\delta) &= \int_0^\delta F_Y(u)du = \delta + (e^{-\lambda\delta} - 1) - \sum_{m=1}^{r-1} \frac{1}{m!} \int_0^\delta (\lambda u)^m e^{-\lambda u}\, du \\ &= \delta + (e^{-\lambda\delta} - 1) - \sum_{m=1}^{r-1} \frac{1}{\lambda}\left[1 - e^{-\lambda\delta} - e^{-\lambda\delta} \sum_{l=1}^{m} (\lambda\delta)^l / l!\right] \end{aligned} \tag{37}$$

and

$$D_m \equiv D_{01m\delta} = \int_{(m-1)\delta}^{m\delta} [1 - F_Y(u)]du = \delta - [J(m\delta) - J((m-1)\delta)] \tag{38}$$

To simplify the notation in the equations below, we will drop the subscripts 0,1,δ on $D_{01m\delta}$ and denote it D_m, and will substitute m_1 for $(m-1)\delta$ and m_2 for $m\delta$. We recall that the mean and variance of h_j are $\mu_j \equiv r_j/\lambda_j$ and $\sigma_j^2 \equiv r_j/\lambda_j^2$.

The form of the bivariate gamma pdf is

$$\begin{aligned} f_{Y,Z}(y,z) &= \phi\gamma(y;r_1,\lambda_1)\gamma(z;r_2,\lambda_2) + (1-\phi)\gamma(y;r_3,\lambda_3)\gamma(z;r_4,\lambda_4) \\ &\equiv \phi h_1(y)h_2(z) + (1-\phi)h_3(y)h_4(z) \end{aligned} \tag{39}$$

where $0 \leq \phi \leq 1$ is a mixing parameter and the univariate gamma pdf is $\gamma(x; r, \lambda) = \lambda^r x^{r-1} e^{-\lambda x}/Gamma(r)$. To simplify the notation, we denote the gamma pdf in (39) $\gamma(\cdot; r_i, \lambda_i)$ as $h_i(\cdot)$, and its mean, r_i/λ_i, by μ_i. Chen, Prorok, and Graf derive the means and variances of Y and Z, as well as their correlation, when $\lambda_1 = \lambda_2$ and $\lambda_3 = \lambda_4$. Under the more general specification (39),

$$E(Y) \equiv \mu_Y = \phi r_1/\lambda_1 + (1-\phi) r_3/\lambda_3 \equiv \phi\mu_1 + (1-\phi)\mu_3 \tag{40}$$

$$E(Z) \equiv \mu_Z = \phi r_2/\lambda_2 + (1-\phi) r_4/\lambda_4 \equiv \phi\mu_2 + (1-\phi)\mu_4 \tag{41}$$

$$Var(Y) \equiv \sigma_Y^2 = \phi(1-\phi)(\mu_1 - \mu_3)^2 + \phi \cdot \mu_1/\lambda_1 + (1-\phi) \cdot \mu_3/\lambda_3 \tag{42}$$

$$Var(Z) \equiv \sigma_Z^2 = \phi(1-\phi)(\mu_2 - \mu_4)^2 + \phi \cdot \mu_2/\lambda_2 + (1-\phi) \cdot \mu_4/\lambda_4 \tag{43}$$

$$E(YZ) = \phi\mu_1\mu_2 + (1-\phi)\mu_3\mu_4 \tag{44}$$

$$Cor(Y, Z) \equiv \rho_{YZ} = \phi(1-\phi)(\mu_1 - \mu_3)(\mu_2 - \mu_4)/(\sigma_Y \sigma_Z) \tag{45}$$

In words, the joint density function is a mixture of a product of two gamma densities: with probability ϕ, $Y \sim Gamma(r_1, \lambda_1)$ and $Z \sim Gamma(r_2, \lambda_2)$; with probability $1-\phi$, $Y \sim Gamma(r_3, \lambda_3)$ and $Z \sim Gamma(r_4, \lambda_4)$. Note that the conditional density of Z given Y is also a mixture of two gamma densities h_2 and h_4, but with mixing proportions α and $1-\alpha$, where $\alpha = (1 + \frac{1-\phi}{\phi}\frac{h_3(y)}{h_1(y)})^{-1}$. When $Y = \mu_1 = r_1/\lambda_1$, α will be close to 1 (i.e., $f_{Z|Y}(z|\mu_1) \approx h_2(z)$), while $Y = \mu_3$ leads to small α (i.e., $f_{Z|Y}(z|\mu_3) \approx h_4(z)$). Also, when $\phi = 0$ or $\phi = 1$, $\mathrm{Cor}(Y, Z) = 0$, and hence the length-biased sampling of Y has no effect on the distribution of Z. Kafadar and Prorok (2001) provide some equations for choosing values of the parameters of the bivariate gamma density so that the marginal means and variances, as well as the correlation, are representative of the values one might expect for sojourn times and clinical durations for cancers of various types (short for pancreatic cancer; moderate for breast cancer; long for prostate cancer).

Because the formulae for $E(Y^*)$ and $E(Z^*)$ and for $Var(Y^*)$ and $Var(Z^*)$ are expressed in terms of $E(Y^*_{(k)})$, $E(Y^*_{(k)})^2$, $E(Z^*_{(k)})$, $E(Z^*_{(k)})^2$, the derivations for these four quantities for the

special case of the bivariate gamma density (39) are given below, first for the special case of (39) as the mixed exponential ($r_i = 1$ for all i = 1,2,3,4), and then for the more general case. Derivations of these formulae are given in the appendix.

Bivariate gamma pdf

When Y and Z have a bivariate gamma density with parameters r_i and λ_i, i = 1,2,3,4 [cf. eqn (39)], the mean of $Y^*_{(k)}$ can be calculated explicitly from:

$$\begin{aligned} f_{Y_k^*}(y) &= g_k(y) f_Y(y) \\ &= \left\{ \beta \sum_{m=1}^{k} (1-\beta)^{m-1} [(y-m_1) I_{(m_1,m_2)}(y) + \delta I_{m_2,\infty}(y)] + \beta(1-\beta)^k (y-k\delta) I_{(k\delta,\infty)}(y) \right\} \cdot \\ &\quad [\phi h_1(y) + (1-\phi) h_3(y)]\,; \;\; m_1 \equiv (m-1)\delta; \; m_2 \equiv m\delta \end{aligned} \tag{46}$$

from which it follows that

$$E(Y_k^*) = \beta \sum_{m=1}^{k} (1-\beta)^{m-1} [(a) - (b) + (c)] + \beta(1-\beta)^k [(d) - (e)] \tag{47}$$

where

$$(a) = \int_{m_1}^{m_2} y^2 [\phi h_1(y) + (1-\phi) h_3(y)] dy \tag{48}$$

$$(b) = (m-1)\delta \int_{m_1}^{m_2} y [\phi h_1(y) + (1-\phi) h_3(y)] dy \tag{49}$$

$$(c) = \delta \int_{m_2}^{\infty} y [\phi h_1(y) + (1-\phi) h_3(y)] dy \tag{50}$$

$$(d) = \int_{k\delta}^{\infty} y^2 [\phi h_1(y) + (1-\phi) h_3(y)] dy \tag{51}$$

$$(e) = k\delta \int_{k\delta}^{\infty} y [\phi h_1(y) + (1-\phi) h_3(y)] dy \tag{52}$$

The formula for $E(Y_k^*)^2$ is similar, except each exponent of y in the above equations is increased by 1. The integrals can be calculated explicitly using Eqn (37) or more simply using the `R` function `pgamma(x,r,lambda)`, from which one can define $P(a,b,r,\lambda) \equiv$ `pgamma(b,r,lambda)` – `pgamma(a,r,lambda)`. Apart from constants, each of these integrals is of the form $\int_a^b y^n h_i(y) dy$,

which equals $P(a,b,r_i,\lambda_i))$, $(r_i/\lambda_i)P(a,b,r_i+1,\lambda_i)$, $(r_i(r_i+1)/\lambda_i^2)P(a,b,r_i+2,\lambda_i)$, $(r_i(r_i+1)(r_i+2)/\lambda_i^3)P(a,b,r_i+3,\lambda_i)$, as $n = 0$, 1, 2, 3, respectively.

Computation of the mean of $Z_k^* = E(g_k(Y)Z)$ is similar; from (35),

$$\begin{aligned} f_{Z_k^*}(z) &= \int_0^\infty \int_0^\infty g_k(y)[\phi h_1(y)h_2(z) + h_3(y)h_4(z)]dydz \\ &= \phi h_2(z) \cdot A + (1-\phi)h_4(z) \cdot B \qquad (53) \\ E(Z_k^*) &= \int_0^\infty z \cdot g_k(y)[\phi h_1(y)h_2(z) + h_3(y)h_4(z)]dydz \\ &= \phi \int_0^\infty g_k(y)h_1(y)dy \int_0^\infty zh_2(z)dz + (1-\phi)\int_0^\infty g_k(y)h_3(y)dy \int_0^\infty zh_4(z)dz \\ &= \phi\mu_2 \int_0^\infty g_k(y)h_1(y)dy + (1-\phi)\mu_4 \int_0^\infty g_k(y)h_3(y)dy \\ &= \phi\mu_2 A + (1-\phi)\mu_4 B \qquad (54) \end{aligned}$$

where $\mu_j \equiv r_j/\lambda_j$, $A \equiv \int_0^\infty g_k(y)h_1(y)dy$, and $B \equiv \int_0^\infty g_k(y)h_3(y)dy$.

5 Illustrations

This section illustrates these calculations for six scenarios, designed to correspond roughly to fast, moderate, and slow durations of both preclinical and clinical disease, as might be seen, for example, with pancreatic cancer, breast cancer, and prostate cancer, respectively. We assume that the test sensitivity, β, is constant for all screens and for all persons (relaxation of this assumption is the subject of current research) and set it to one of four values: 0.70, 0.80, 0.90, 0.95. (Screening tests with test sensitivities below 0.70 usually are deemed too impractical for general screening, while $\beta = 0.95$ would be considered high sensitive.) We also perform the calculations when the screening interval is 1 year, 2 years, and 3 years; as the screening interval increases, one expects to see an increase in the effect of the length-biased sampling. Finally, we consider three different mixing proportions $\phi = 0.3, 0.5, 0.7$, for each specification of $f_{YZ}(y,z)$ (cf. eqn (39)), yielding a total of 18 joint pdfs (six scenarios $\times$ three values of ϕ).

The six scenarios are:

(A): Combination of fast and slow growth ($\rho = 0.57, 0.66, 0.67$):

Preclinical duration: $\phi\Gamma(3,2) + (1-\phi)\Gamma(10,2)$ (means 1.5, 5)

Clinical duration: $\phi\Gamma(4,2) + (1-\phi)\Gamma(12,2)$ (means 2, 6)

(B): Combination of fast and moderate growth ($\rho = 0.23, 0.29, 0.28$):

Preclinical duration: $\phi\Gamma(2,2) + (1-\phi)\Gamma(4,2)$ (means 1, 2)

Clinical duration: $\phi\Gamma(3,2) + (1-\phi)\Gamma(6,2)$ (means 1.5, 3)

(C): Combination of moderate and slow growth ($\rho = 0.63, 0.71, 0.72$):

Preclinical duration: $\phi\Gamma(6,2) + (1-\phi)\Gamma(14,2)$ (means 3, 7)

Clinical duration: $\phi\Gamma(4,2) + (1-\phi)\Gamma(18,2)$ (means 2, 9)

(D): Combination of speedy and fast growth ($\rho = 0.12, 0.15, 0.14$):

Preclinical duration: $\phi\Gamma(2,2) + (1-\phi)\Gamma(3,2)$ (means 1, 1.5)

Clinical duration: $\phi\Gamma(2,2) + (1-\phi)\Gamma(4,2)$ (means 1, 2)

(E): Combination of speedy and moderate growth ($\rho = 0.39, 0.50, 0.56$):

Preclinical duration: $\phi\Gamma(2,4) + (1-\phi)\Gamma(3,1)$ (means 0.5, 3)

Clinical duration: $\phi\Gamma(2,4) + (1-\phi)\Gamma(4,2)$ (means 0.5, 2)

(F): Combination of slow and very slow growth ($\rho = 0.60, 0.67, 0.66$):

Preclinical duration: $\phi\Gamma(16,2) + (1-\phi)\Gamma(28,2)$ (means 8, 14)

Clinical duration: $\phi\Gamma(8,2) + (1-\phi)\Gamma(20,2)$ (means 4, 10)

The three values for the correlation ρ arise from three different values of ϕ (0.3, 0.5, 0.7), for each scenario, permitting a total of 18 bivariate pdfs for the joint density of Y and Z. Table 1 lists the characteristics of each pdf (parameters of $h_i(\cdot)$, $i = 1, 2, 3, 4$; ϕ; mean and standard deviation of

Y and of Z; correlation between Y and Z). The last two columns of this table give the weight on h_2 given $Y = \mu_1$ or $Y = \mu_3$; i.e., $f_{Z|Y}(z|y = \mu_i) = \alpha(\mu_i)h_2(z) + (1 - \alpha(\mu_i))h_4(z)$, where $\alpha(y) = (1 + \frac{1-\phi}{\phi}\frac{h_3(y)}{h_1(y)})^{-1}$.

Figures 3 through 7 show the pdf of Y (left column), the pdf of Z (middle column), and the joint pdf $f_{YZ}(y, z)$ (right column), for three values of ρ: 0.3 (top row), 0.5 (middle row), 0.7 (bottom row). Notice that the bimodality of Y and Z is hardly apparent in Scenarios B and D, and that the value of ϕ affects the relative proportion of the two components in the mixture of the univariate and bivariate pdfs.

Table 1

pdf	r_1	r_2	r_3	r_4	ϕ	E(Y)	SD(Y)	E(Z)	SD(Z)	Corr	$\alpha(\mu_1)$	$\alpha(\mu_3)$
A1	3	4	10	12	0.3	3.95	2.1325	4.80	2.4000	0.5744	0.9726	0.0077
A2	3	4	10	12	0.5	3.25	2.1651	4.00	2.4495	0.6600	0.9881	0.0178
A3	3	4	10	12	0.7	2.55	1.9615	3.20	2.2271	0.6730	0.9949	0.0406
B1	2	3	4	6	0.3	1.70	1.0296	2.55	1.3219	0.2314	0.3913	0.1385
B2	2	3	4	6	0.5	1.50	1.0000	2.25	1.2990	0.2887	0.6000	0.2727
B3	2	3	4	6	0.7	1.30	0.9274	1.95	1.2031	0.2823	0.7778	0.4667
C1	6	4	14	18	0.3	5.80	2.5020	6.90	3.7068	0.6340	0.9298	0.0148
C2	6	4	14	18	0.5	5.00	2.5495	5.50	3.8730	0.7089	0.9686	0.0340
C3	6	4	14	18	0.7	4.20	2.3367	4.10	3.5128	0.7163	0.9863	0.0758
D1	2	2	3	4	0.3	1.35	0.8529	1.70	1.0296	0.1196	0.3000	0.2222
D2	2	2	3	4	0.5	1.25	0.8292	1.50	1.0000	0.1508	0.5000	0.4000
D3	2	2	3	4	0.7	1.15	0.7921	1.30	0.9274	0.1429	0.7000	0.6087
E1	2	2	3	4	0.3	2.25	1.8574	1.55	1.1000	0.3854	0.8596	0.0006
E2	2	2	3	4	0.5	1.75	1.7678	1.25	1.0607	0.5000	0.9346	0.0013
E3	2	2	3	4	0.7	1.25	1.5166	0.95	0.9274	0.5599	0.9709	0.0031
F1	16	8	28	20	0.3	12.20	3.6959	8.20	3.4147	0.5990	0.9269	0.0151
F2	16	8	28	20	0.5	11.00	3.8079	7.00	3.5355	0.6685	0.9673	0.0346
F3	16	8	28	20	0.7	9.80	3.5299	5.80	3.2342	0.6622	0.9857	0.0772

Note: $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 2$ for all six situations, *except* Situation E, where $\lambda_1 = \lambda_2 = 4$ and $\lambda_3 = 1$, $\lambda_4 = 2$. The last two columns give the proportion, $\alpha(y)$, of $h_2(z)$ (versus $h_4(z)$) when

$y = \mu_1$ or $y = \mu_3$, the mean of $h_1(y)$ or of $h_3(y)$.

Figures 9 through 14 show the relative increase in the means, $E(Y^*)/E(Y)$ and $E(Z^*)/E(Z)$ (left column) and the standard deviations, $SD(Y^*)/SD(Y)$ and $SD(Z^*)/SD(Z)$ (right column), as a function of ϕ. The plot character (1,2,3) denotes the value of δ, and the color (black, red, green blue) denotes the value of β (0.70, 0.80, 0.90, 0.95). As a function of ϕ, the increase tends to be a linear for the mean and slightly quadratic for the standard deviation. Figures 15 through 20 show the same relative increases, but where now the plot character (1,2,3,4) denotes the value of β and the color (black, red, green) denotes the value of δ, along with the fitted lines and quadratics.

These figures show that the increase in the mean and standard deviation of Y^* and Z^* due to the length-biased sampling effect can be substantial: over these six scenarios, test sensitivity 0.70–0.95, and with a prevalent and five period screens,

6 Analyses

Figures 9–14 suggest that ϕ, the proportion of fast versus slow-growing disease, may have the largest effect on the mean and standard deviation of Y^* and Z^*. Table 3 shows the values of the F-statistics in an analysis of veriance of $E(Y^*)$, $SD(Y^*)$, $E(Z^*)$, $SD(Z^*)$, on ϕ, β, δ, $\phi \times \beta$, $\phi \times \delta$, and $\beta \times \delta$, for each of the six scenarios. In most scenarios, ϕ is the dominant source of variation, followed by δ and the interaction $\phi \times \delta$. (I didn't have time to put the table into the text. I also did further analyses on the intercepts and slopes in previous figures that I will show you.)

From the equations for the densities of Y^* and Z^*, we can plot them alongside the underlying densities of Y and Z. These densities are shown in Figures **??**–**??**; the top row shows $f_Y(\cdot)$ as a solid (black) line and $f_{Y^*}(\cdot)$ as a dashed (red) line; the bottom row shows $f_Z(\cdot)$ and $f_{Y^*}(\cdot)$; the three columns correspond to $\phi = 0.3$, 0.5, 0.7. The difference between the two density functions is large especially for $\beta = 0.7$ and $\delta = 3$, Interestingly, it is negligible for Scenario F, where both the

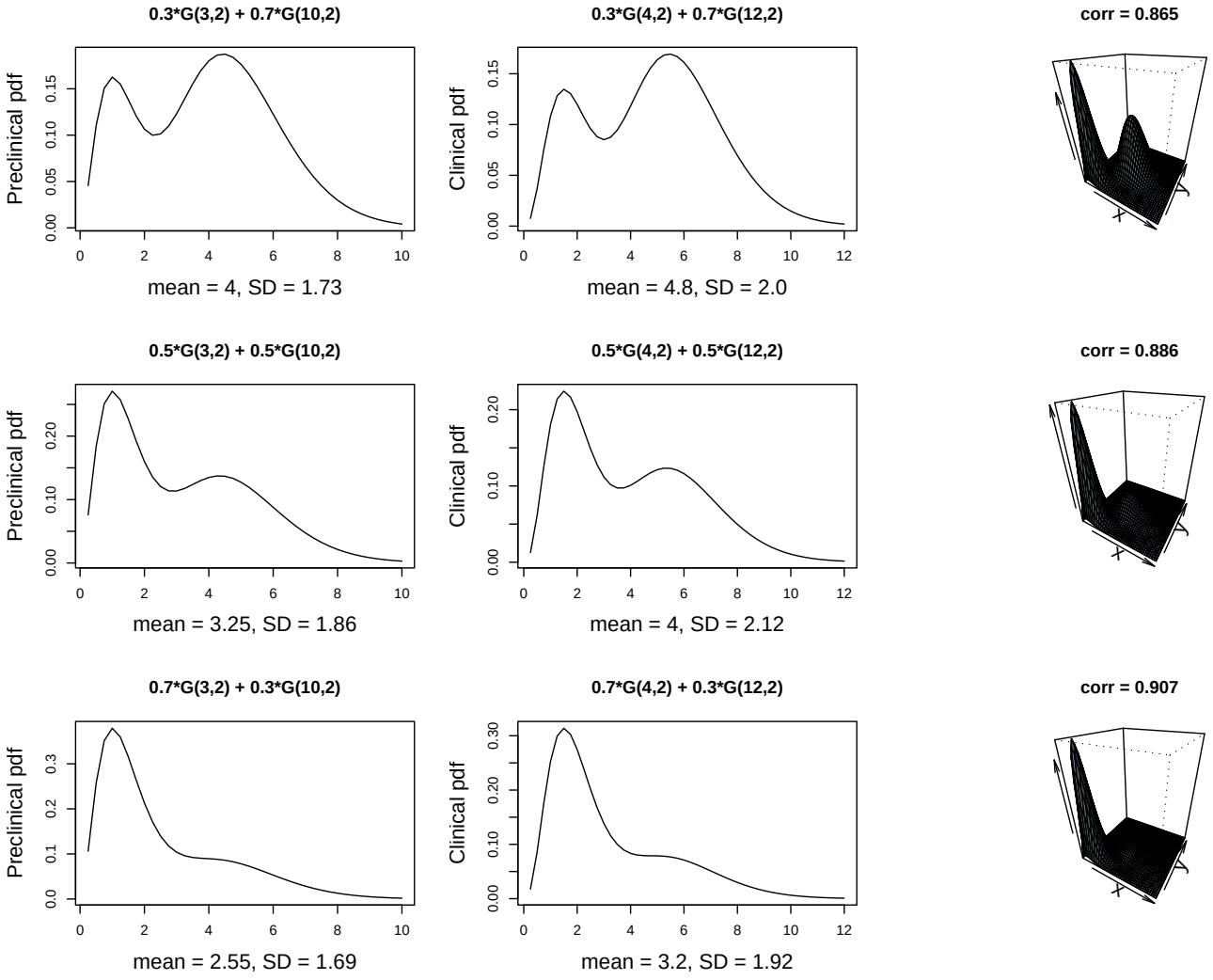

Figure 3: Scenario A: $f_Y(\cdot)$ (left column), $f_Z(\cdot)$ (middle column), $f_{YZ}(\cdot,\cdot)$ (right column); for 3 values of ϕ: 0.3 (top row), 0.5 (middle row), 0.7 (bottom row). See Eqn (39) for $f_{YZ}(\cdot,\cdot)$.

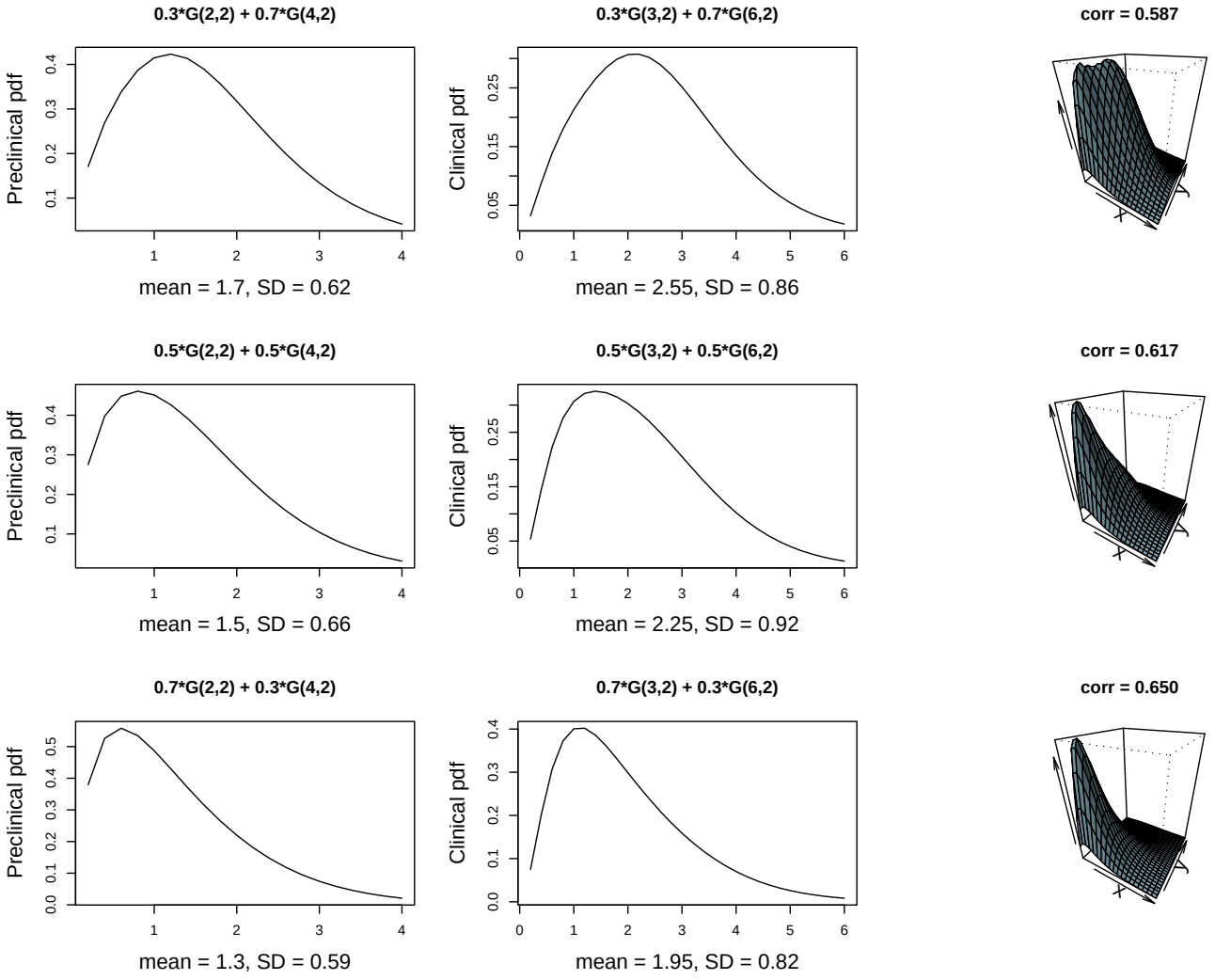

Figure 4: Scenario B: $f_Y(\cdot)$ (left column), $f_Z(\cdot)$ (middle column), $f_{YZ}(\cdot,\cdot)$ (right column); for 3 values of ϕ: 0.3 (top row), 0.5 (middle row), 0.7 (bottom row). See Eqn (39) for $f_{YZ}(\cdot,\cdot)$.

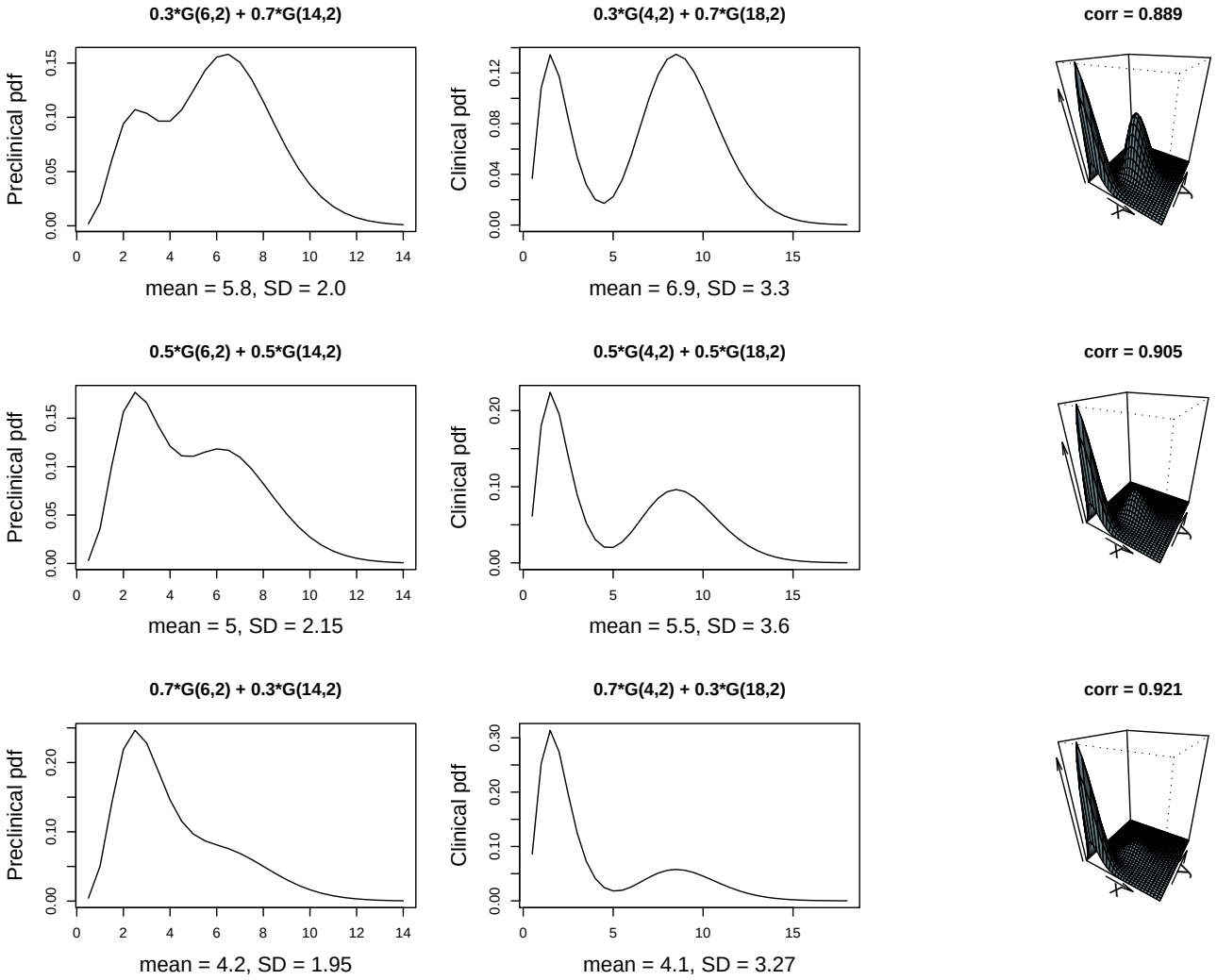

Figure 5: Scenario C: $f_Y(\cdot)$ (left column), $f_Z(\cdot)$ (middle column), $f_{YZ}(\cdot,\cdot)$ (right column); for 3 values of ϕ: 0.3 (top row), 0.5 (middle row), 0.7 (bottom row). See Eqn (39) for $f_{YZ}(\cdot,\cdot)$.

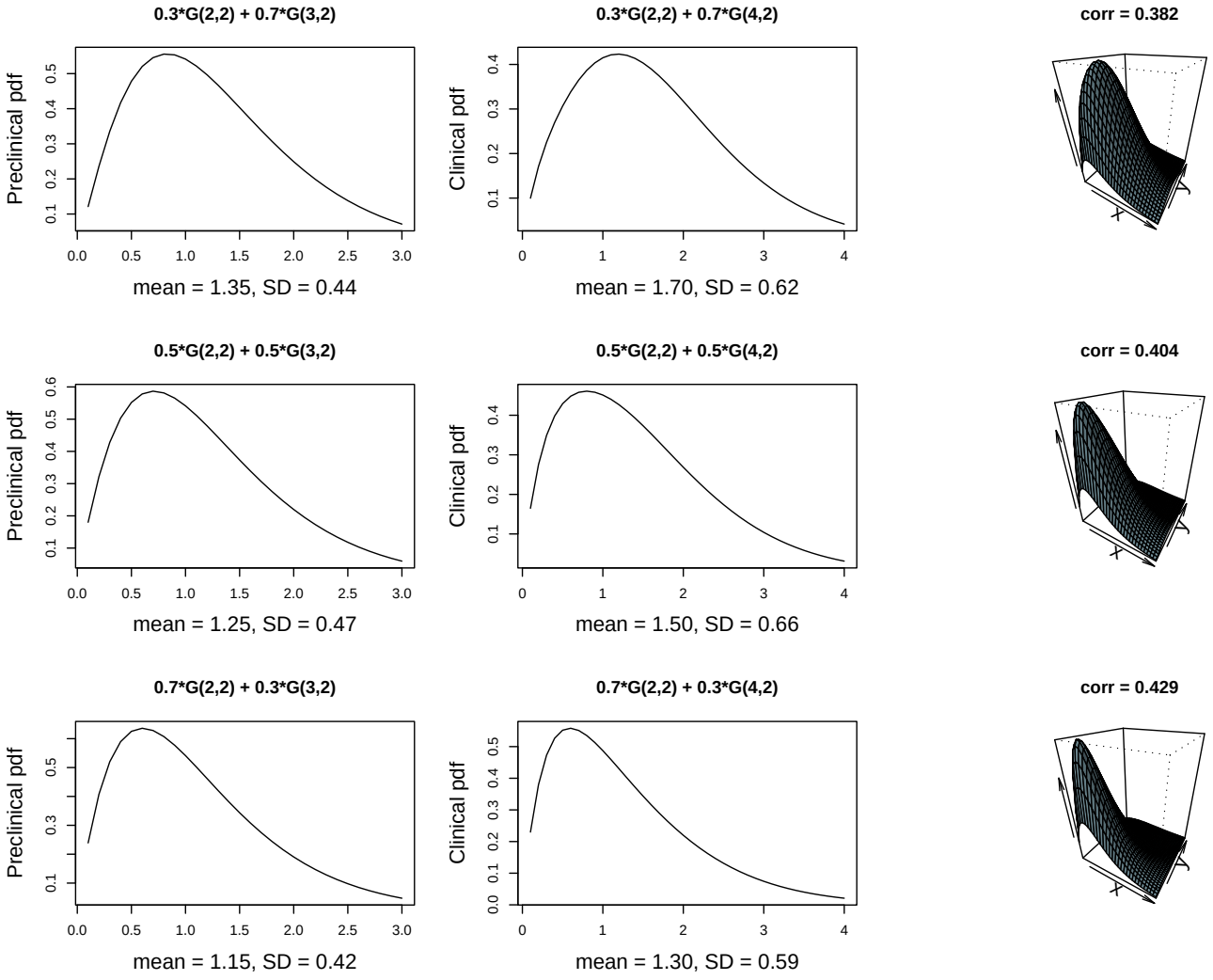

Figure 6: Scenario D: $f_Y(\cdot)$ (left column), $f_Z(\cdot)$ (middle column), $f_{YZ}(\cdot,\cdot)$ (right column); for 3 values of ϕ: 0.3 (top row), 0.5 (middle row), 0.7 (bottom row). See Eqn (39) for $f_{YZ}(\cdot,\cdot)$.

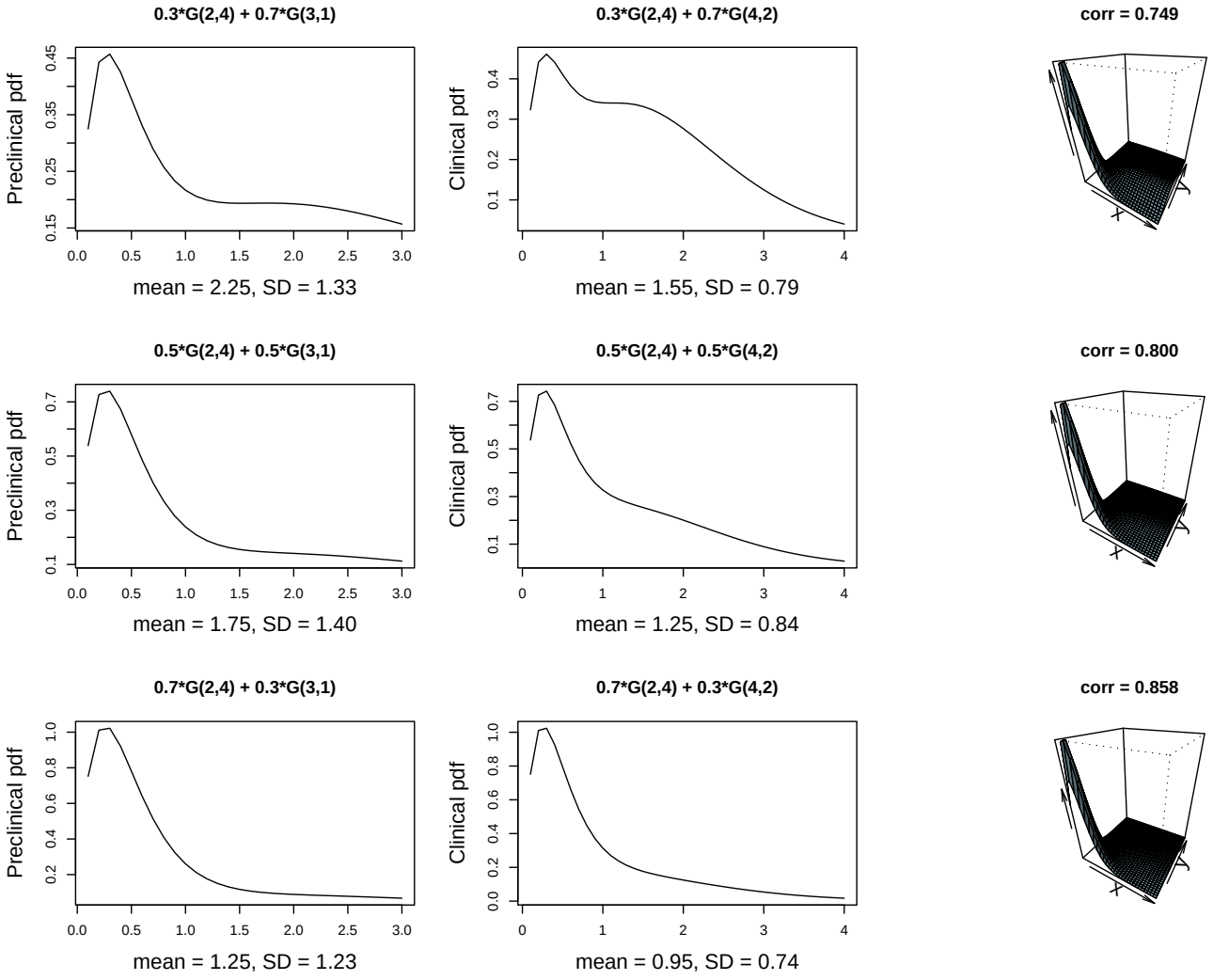

Figure 7: Scenario E: $f_Y(\cdot)$ (left column), $f_Z(\cdot)$ (middle column), $f_{YZ}(\cdot,\cdot)$ (right column); for 3 values of ϕ: 0.3 (top row), 0.5 (middle row), 0.7 (bottom row). See Eqn (39) for $f_{YZ}(\cdot,\cdot)$.

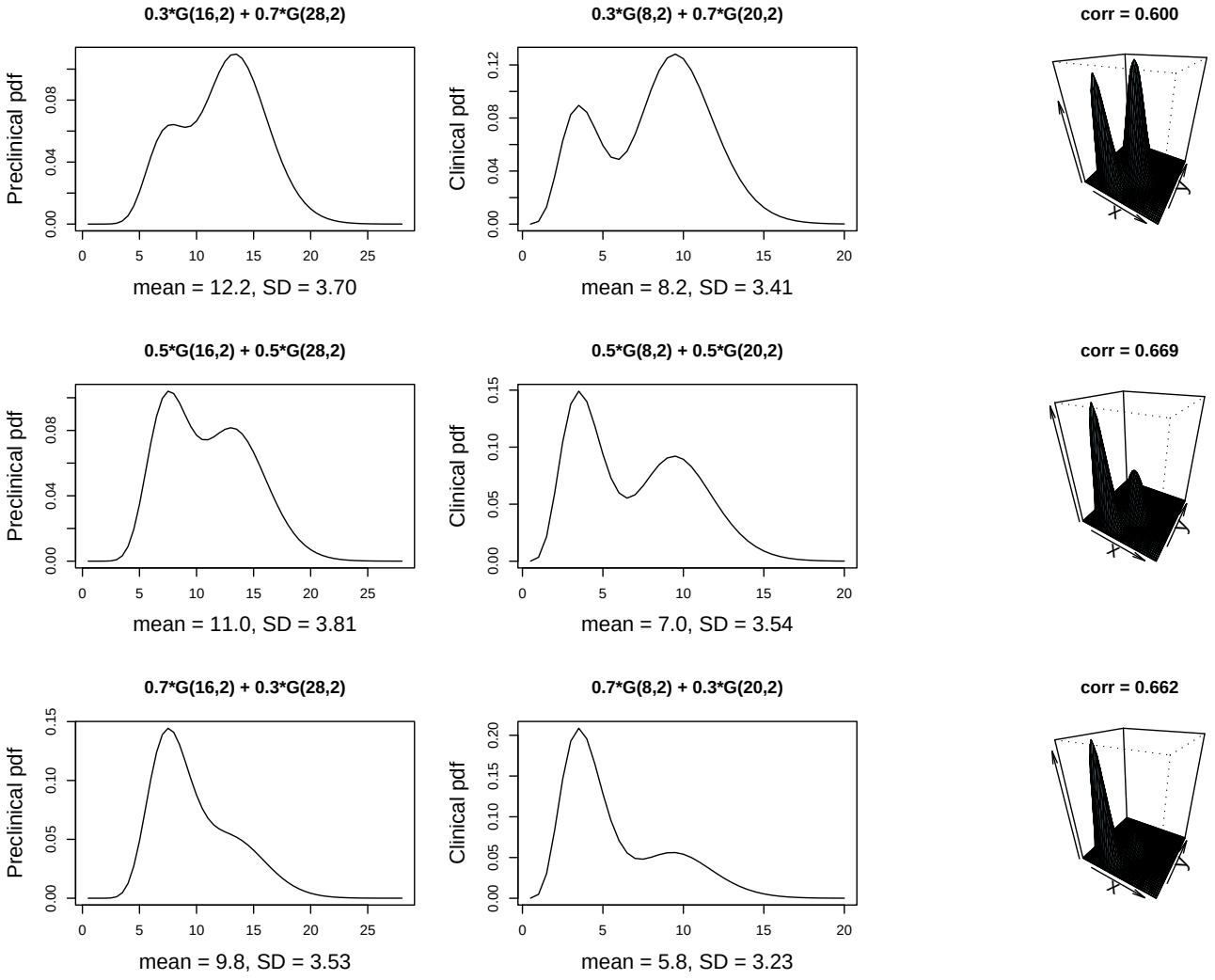

Figure 8: Scenario F: $f_Y(\cdot)$ (left column), $f_Z(\cdot)$ (middle column), $f_{YZ}(\cdot,\cdot)$ (right column); for 3 values of ϕ: 0.3 (top row), 0.5 (middle row), 0.7 (bottom row). See Eqn (39) for $f_{YZ}(\cdot,\cdot)$.

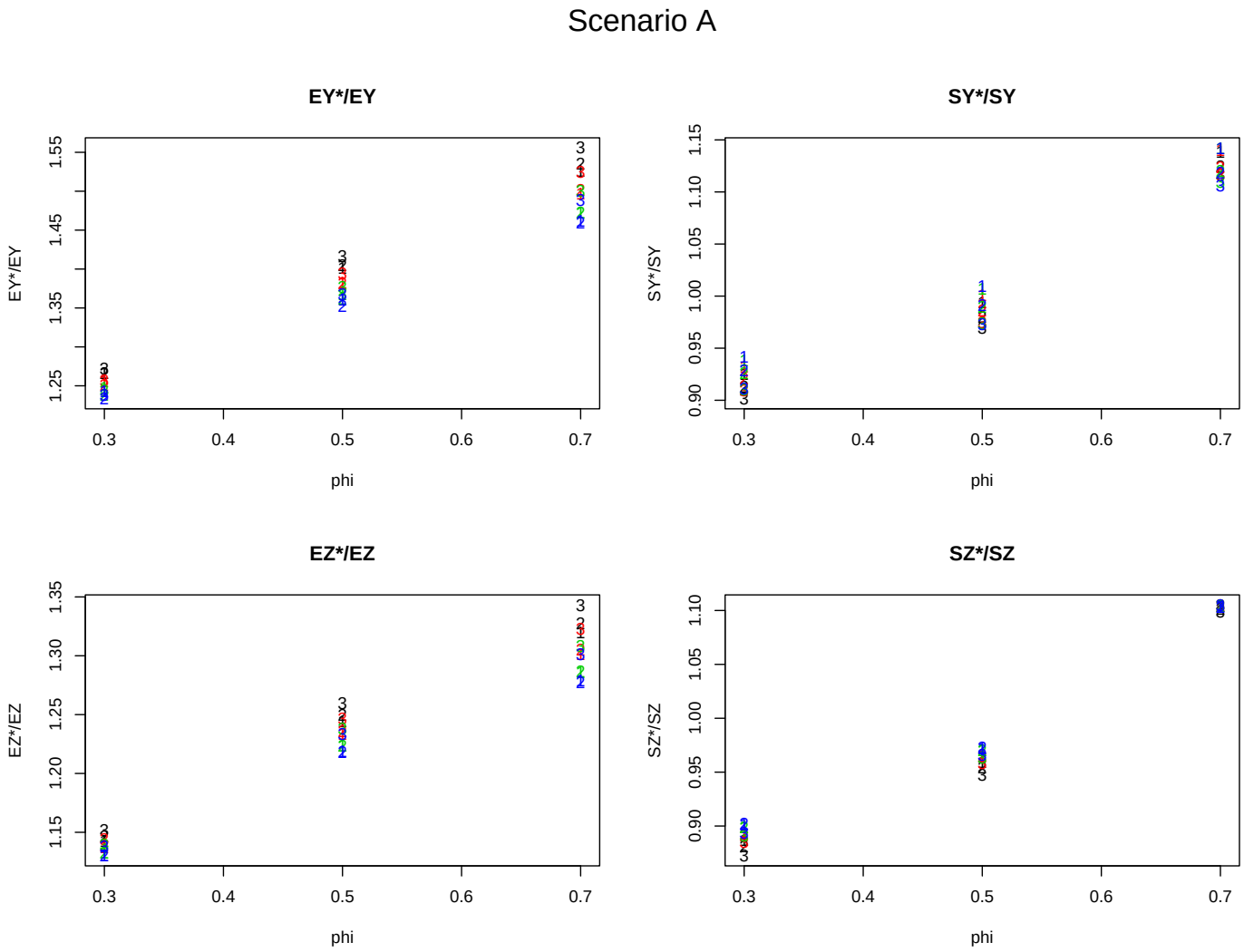

Figure 9: Scenario A: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

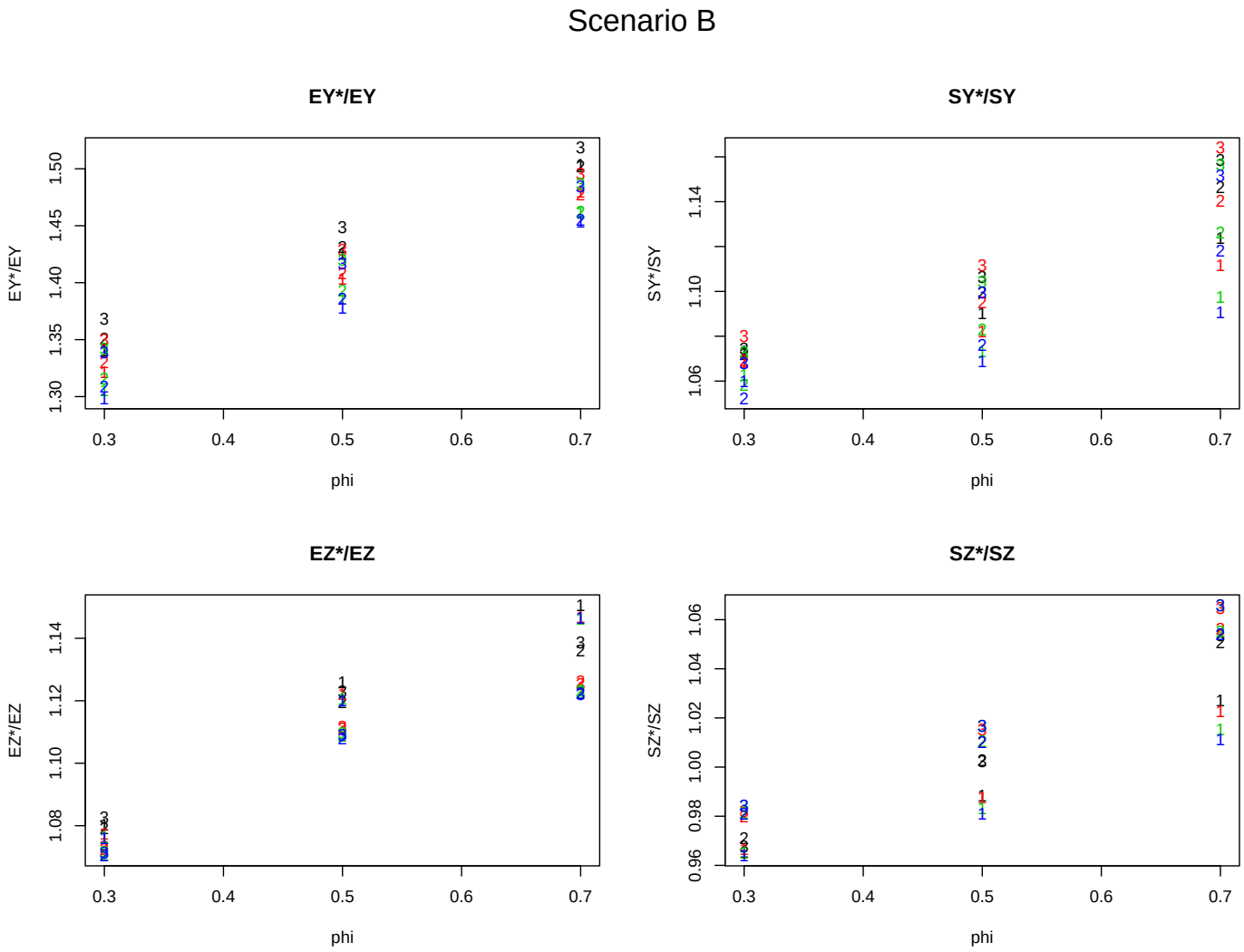

Figure 10: Scenario B: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

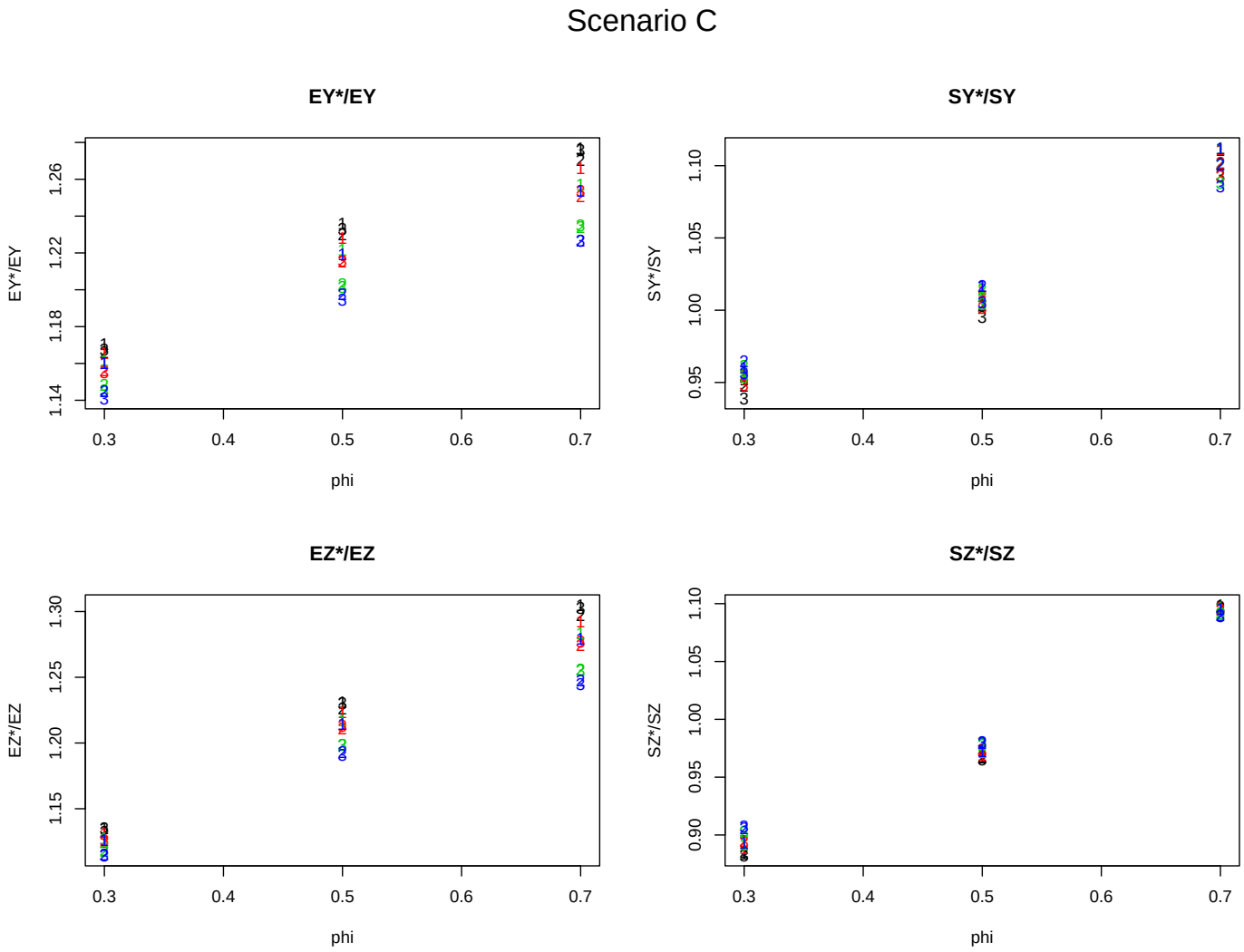

Figure 11: Scenario C: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

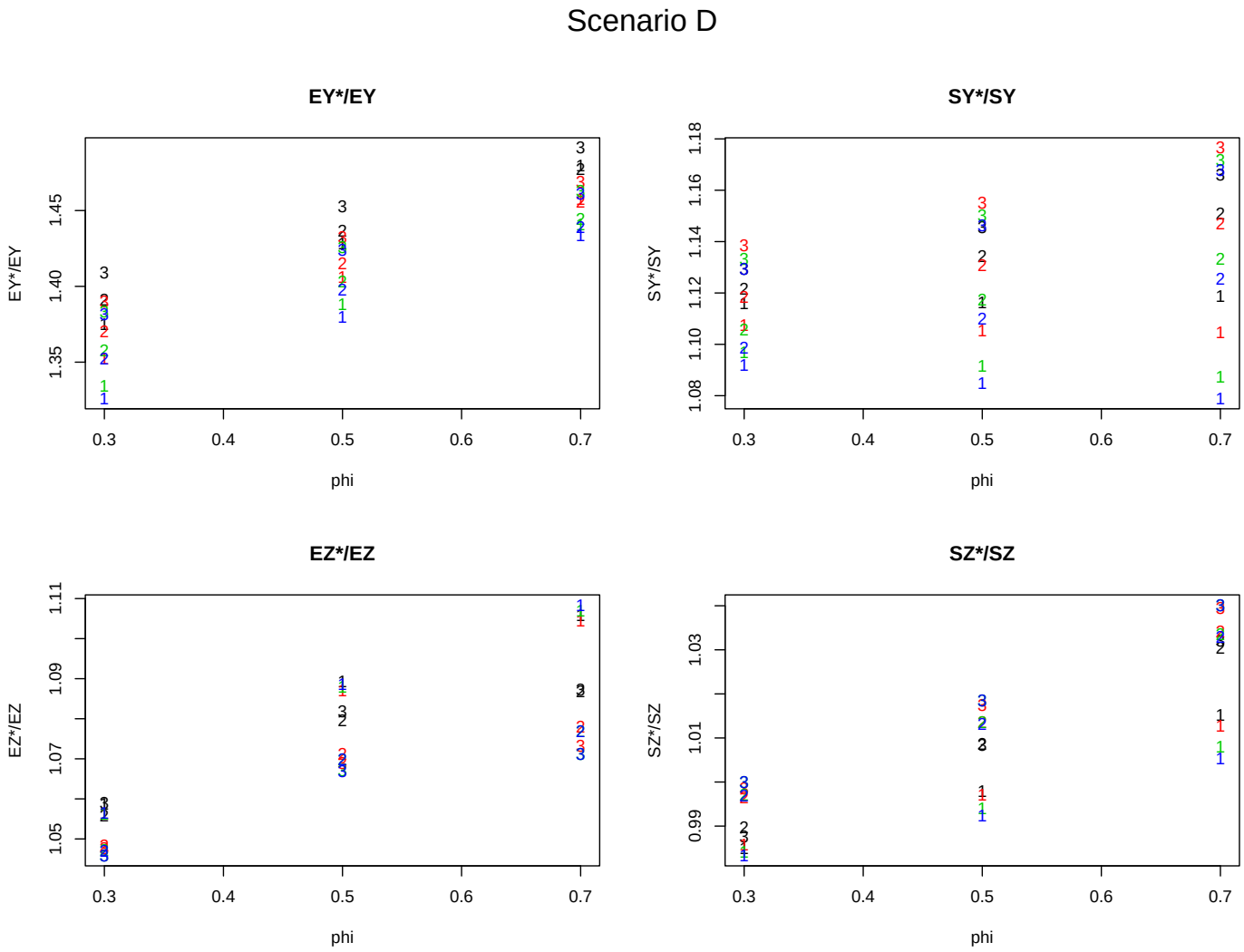

Figure 12: Scenario D: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

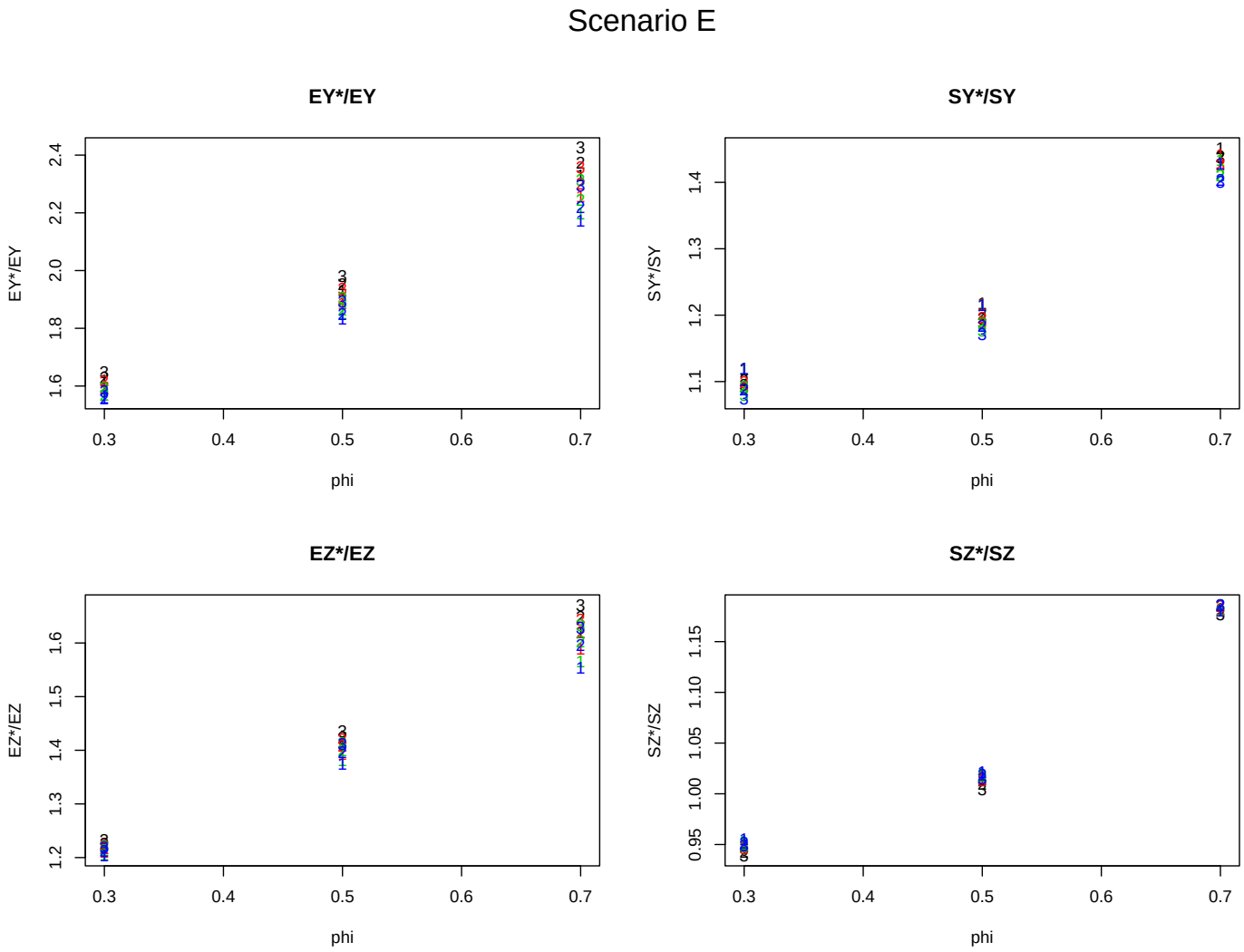

Figure 13: Scenario E: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

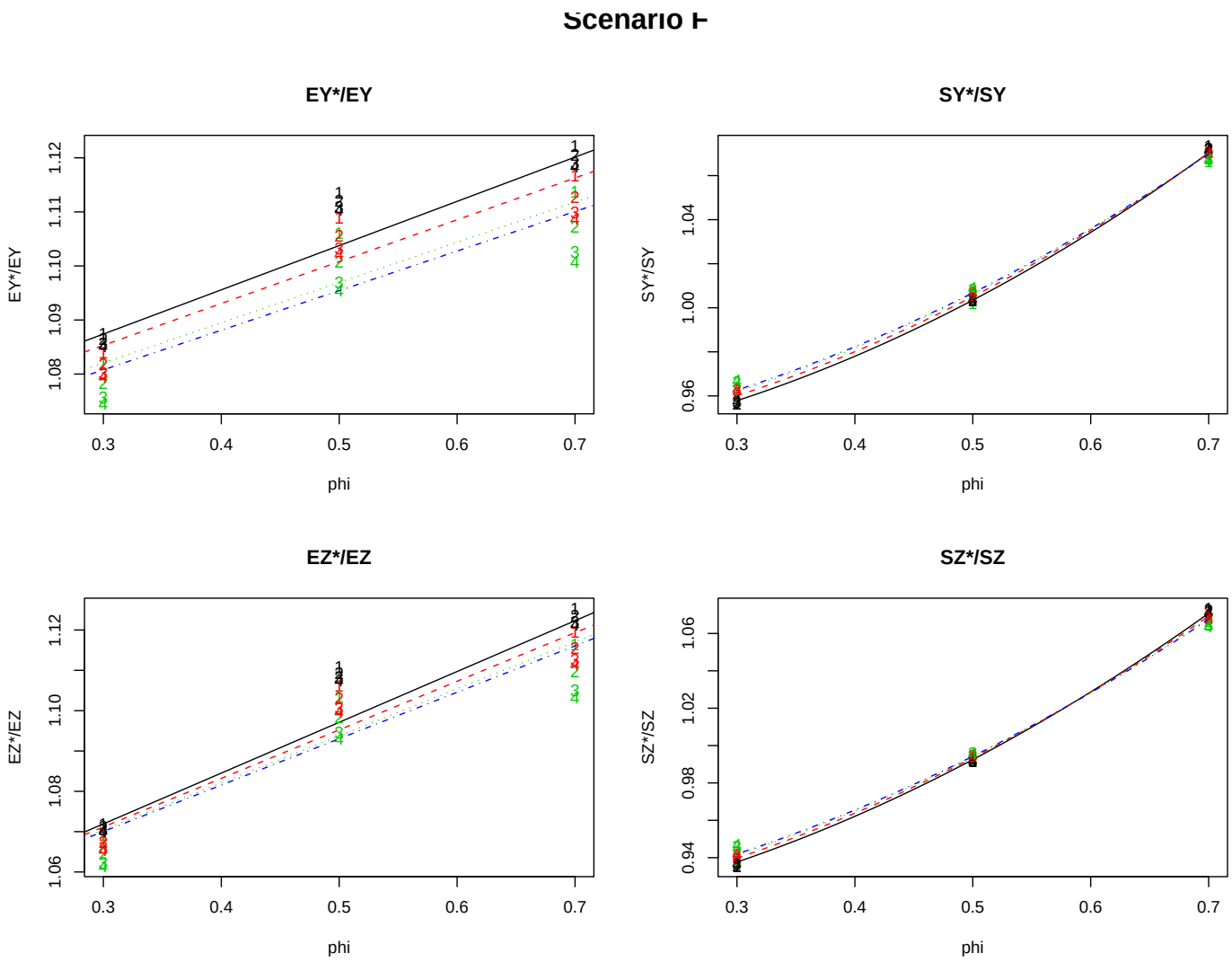

Figure 14: Scenario F: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

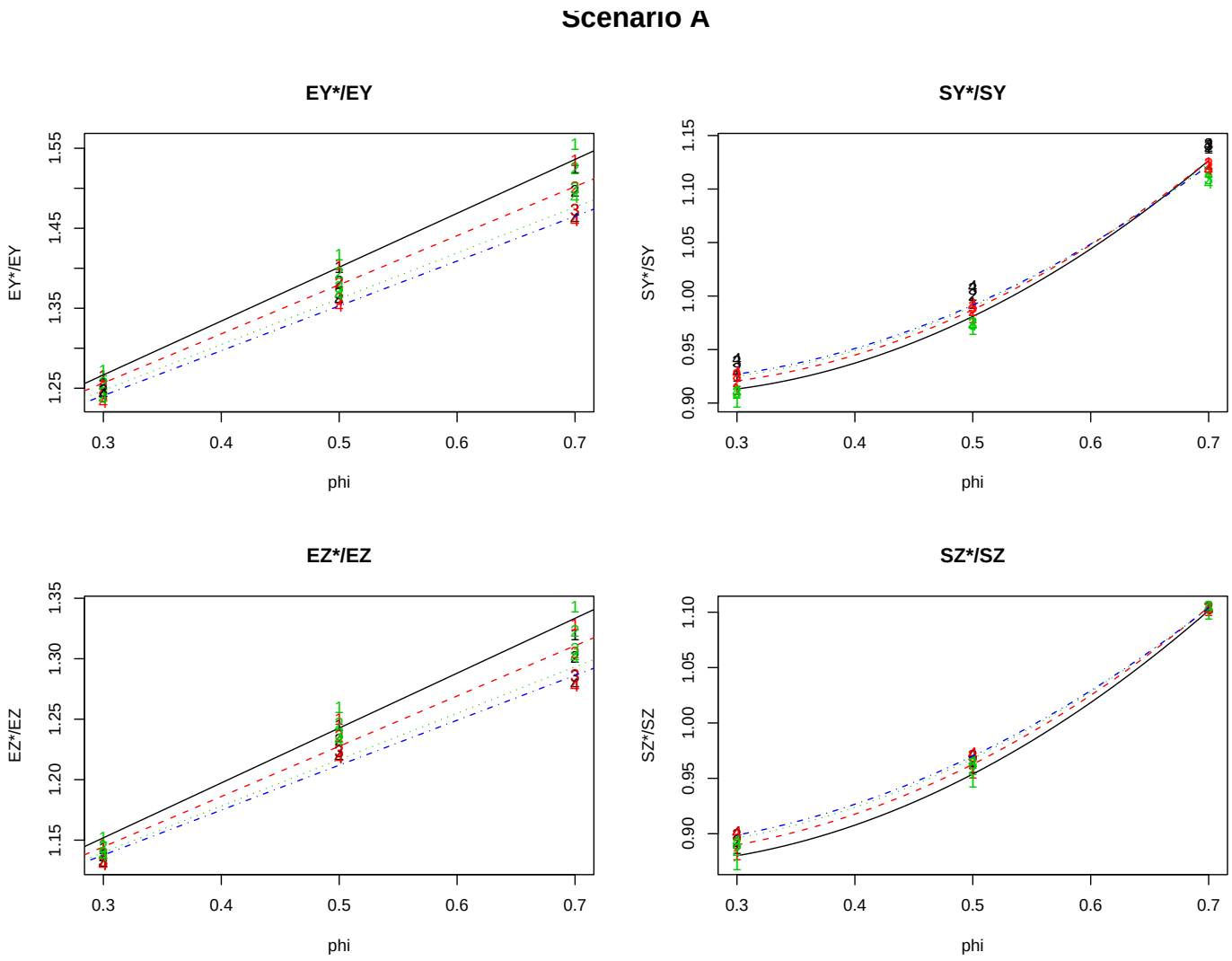

Figure 15: Scenario A: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Fitted line or quadratic for β (test sensitivity) 0.70 (solid black), 0.80 (dash red), 0.90 (dot green), 0.95 (dot-dash blue). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

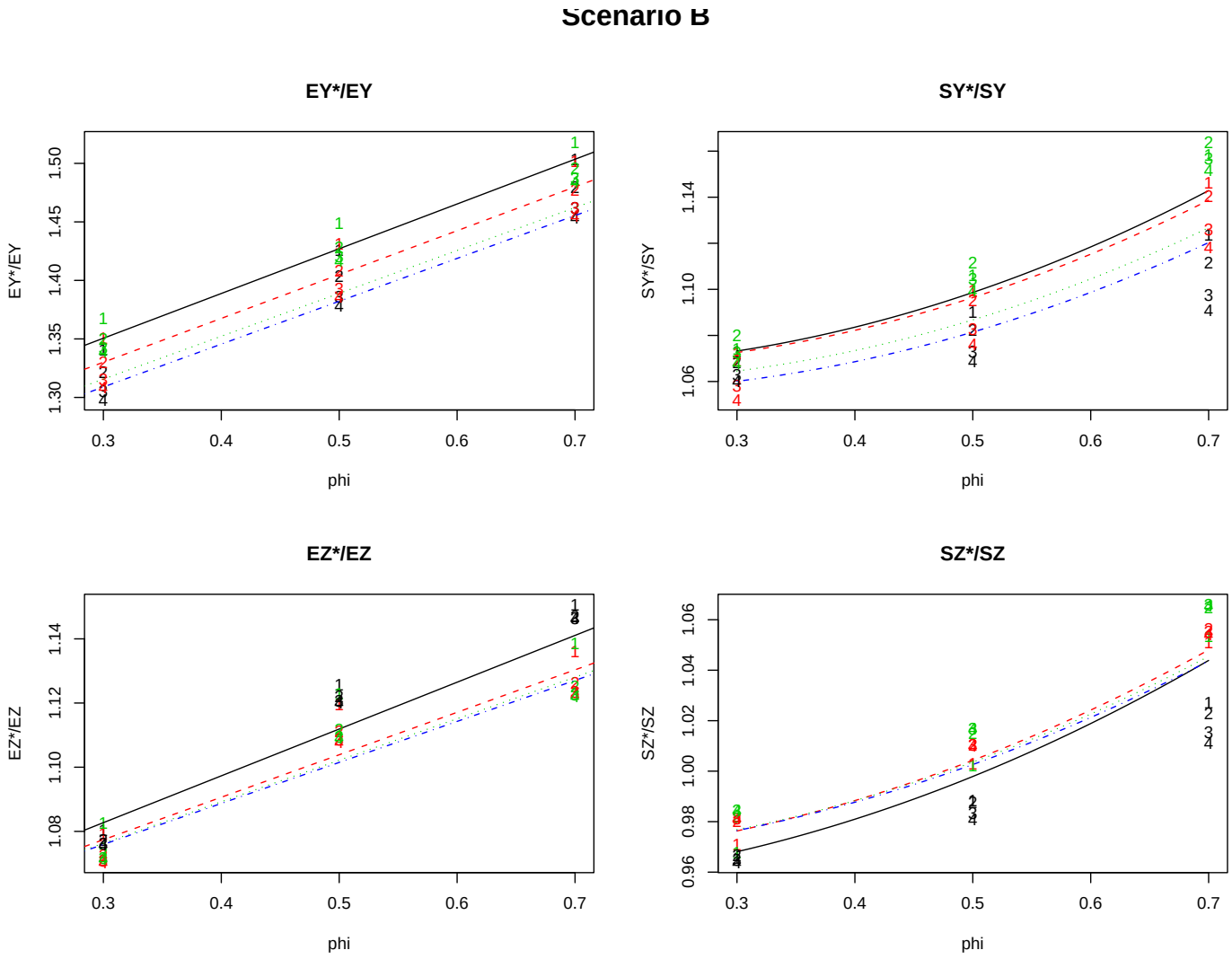

Figure 16: Scenario B: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Fitted line or quadratic for β (test sensitivity) 0.70 (solid black), 0.80 (dash red), 0.90 (dot green), 0.95 (dot-dash blue). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

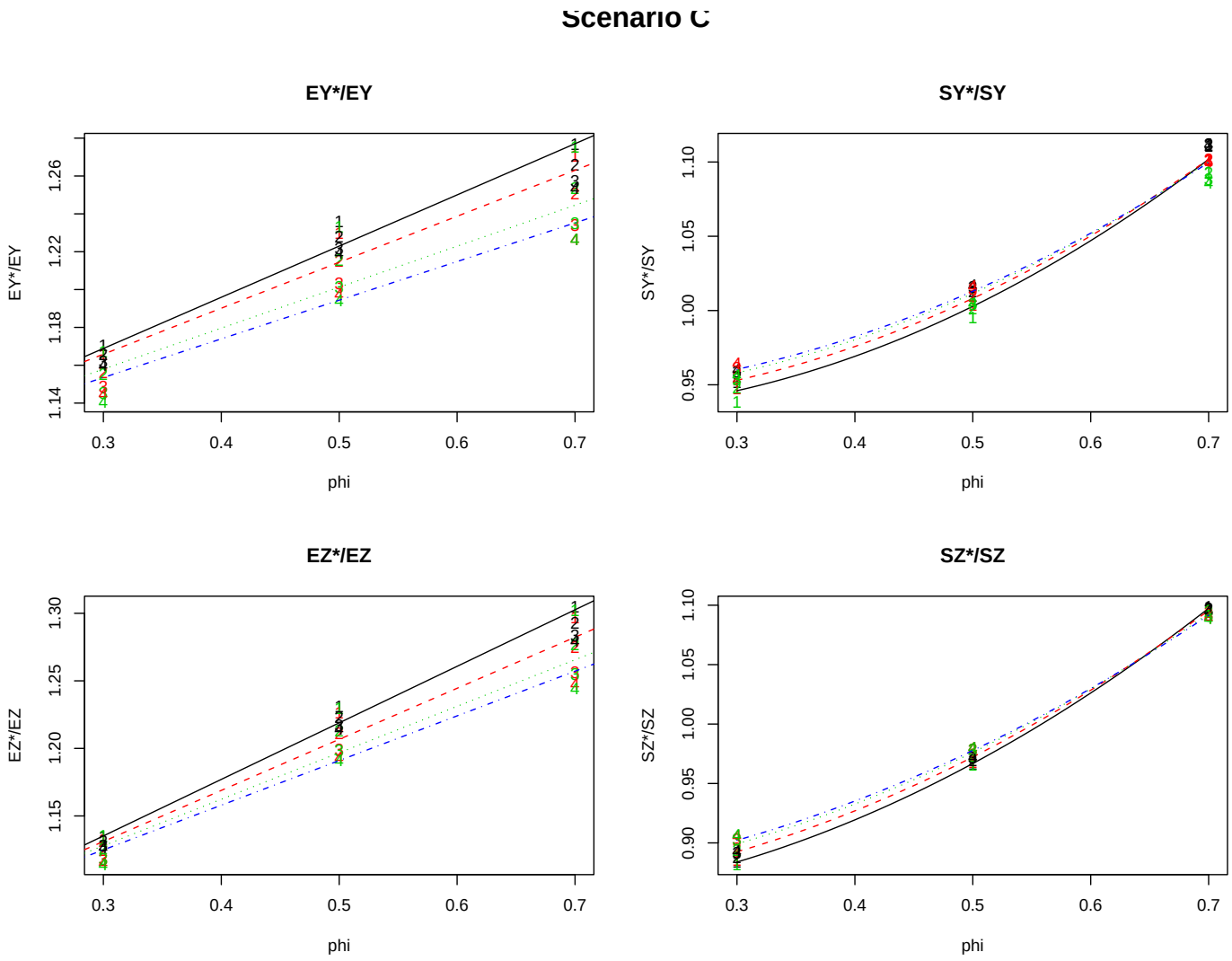

Figure 17: Scenario C: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Fitted line or quadratic for β (test sensitivity) 0.70 (solid black), 0.80 (dash red), 0.90 (dot green), 0.95 (dot-dash blue). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

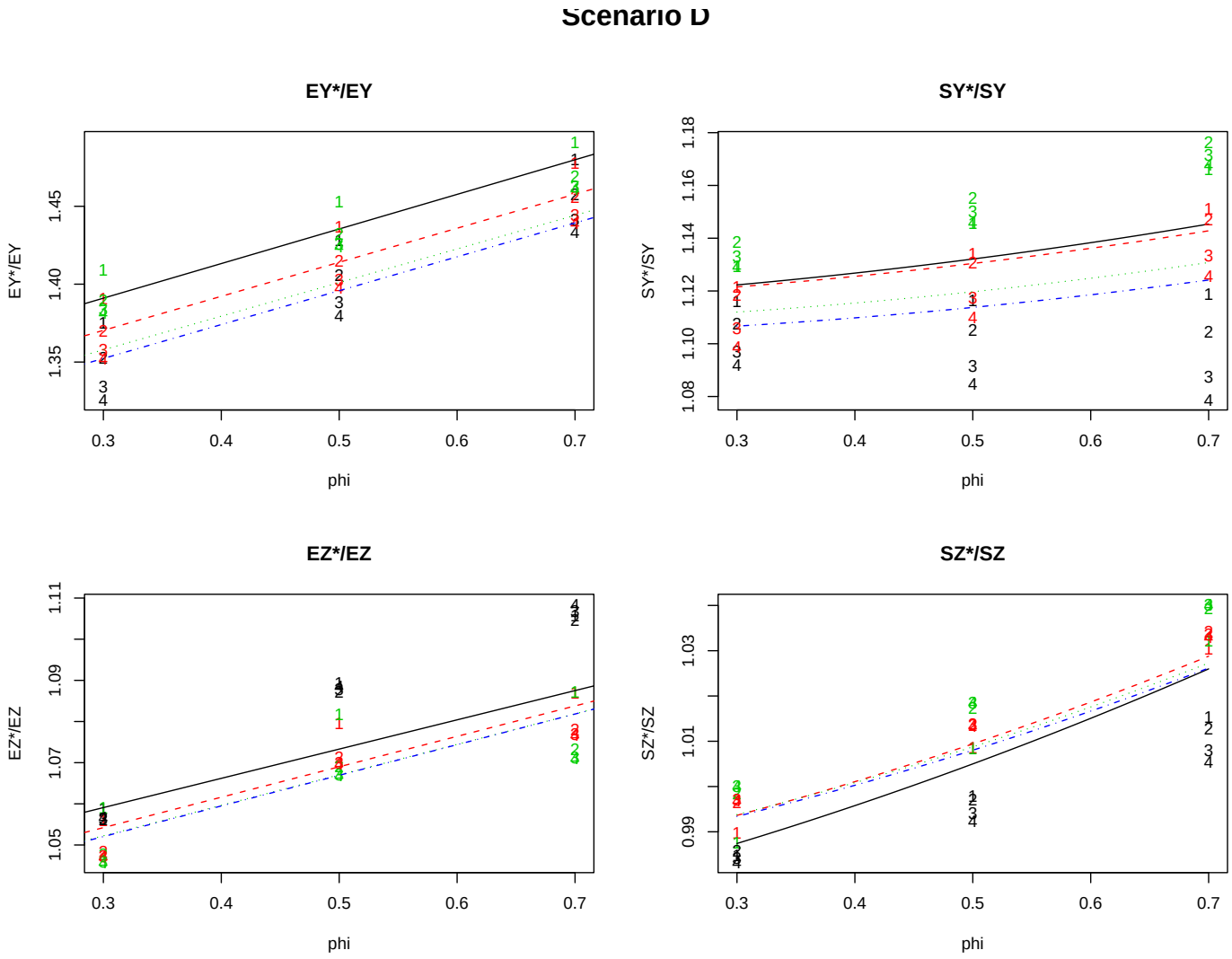

Figure 18: Scenario D: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Fitted line or quadratic for β (test sensitivity) 0.70 (solid black), 0.80 (dash red), 0.90 (dot green), 0.95 (dot-dash blue). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

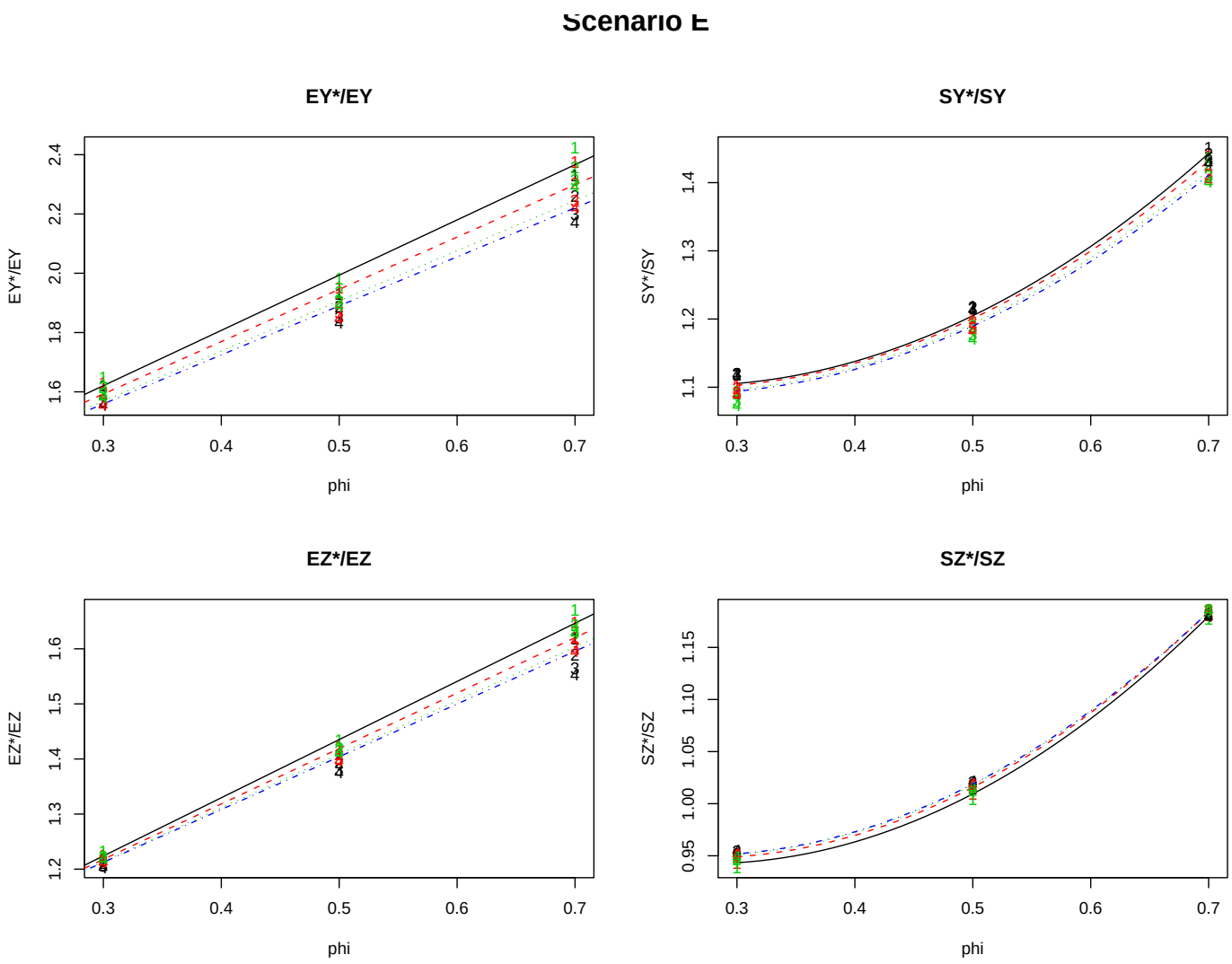

Figure 19: Scenario E: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Fitted line or quadratic for β (test sensitivity) 0.70 (solid black), 0.80 (dash red), 0.90 (dot green), 0.95 (dot-dash blue). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

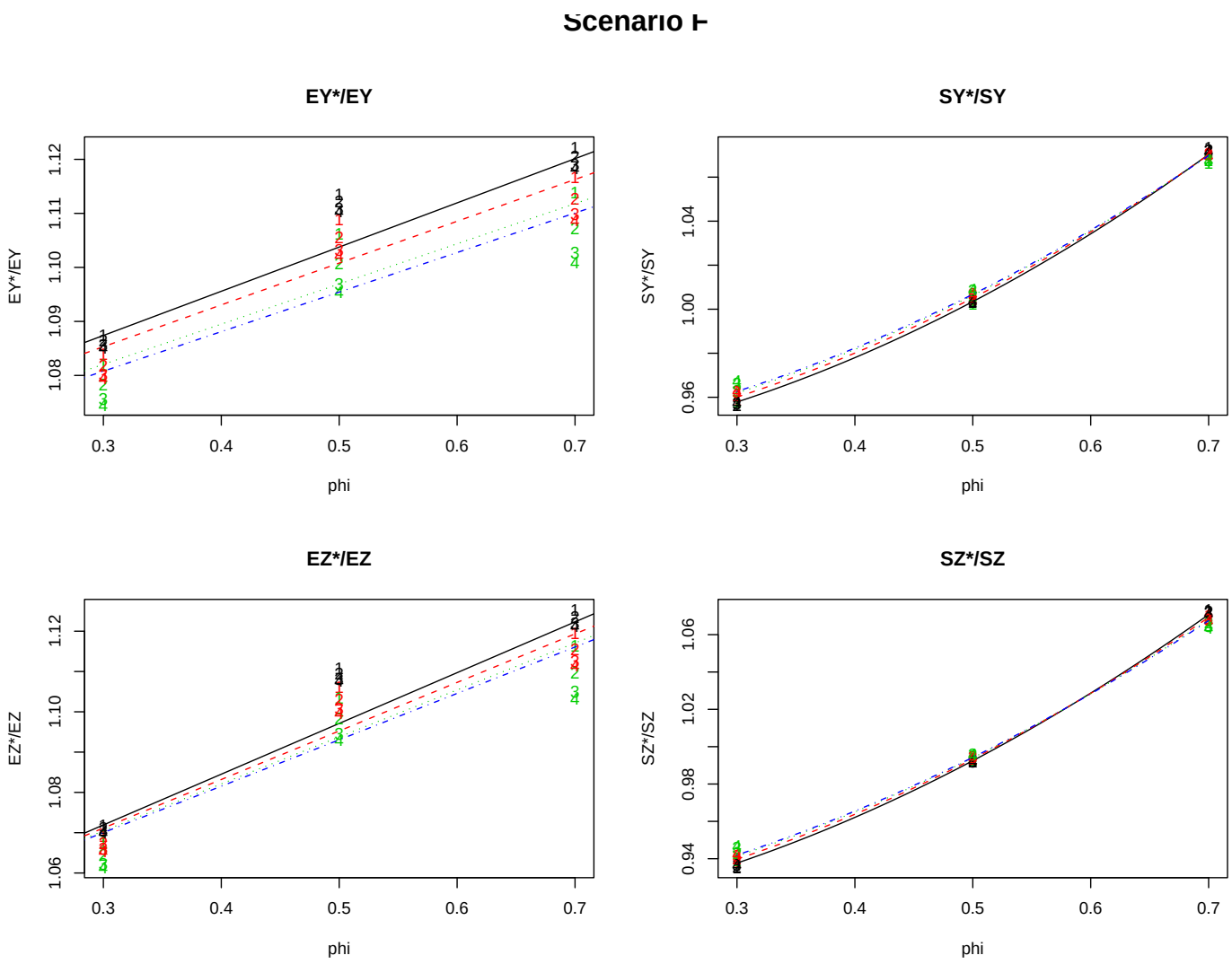

Figure 20: Scenario F: Ratio of increase in mean and standard deviation of Y (sojourn time) and of Z (clinical duration), due to length biased sampling of Y resulting from a prevalent and $k = 5$ subsequent screens, as a function of ϕ, the proportion of short-duration disease (see (39)). Fitted line or quadratic for β (test sensitivity) 0.70 (solid black), 0.80 (dash red), 0.90 (dot green), 0.95 (dot-dash blue). Character value (1, 2, 3) indicates value of δ, screening frequency. (a) $E(Y_k^*)/E(Y)$, as linear function of ϕ. (b) $Var(Y_k^*)/Var(Y)$, as quadratic function of ϕ. (c) $E(Z_k^*)/E(Z)$, as linear function of ϕ. (d) $Var(Z_k^*)/Var(Z)$, as quadratic function of ϕ.

"slow" and "long" disease are still relatively long compared to the screening frequency. (Phil: we will look at the 6 x 3 x 4 plots when you are here. The file is huge and it couldn't be uploaded in time to send to you.)

Finally, the table below shows the maximum increase in the 90^{th} percentiles in Y^* and Z^*, relative to 90^{th} percentiles in Y and Z, $q_{0.90}(Y^*)/q_{0.90}(Y)$ $q_{0.90}(Z^*)/q_{0.90}(Z)$ across the 6 scenarios for different values of β and δ. The increase in these percentiles can be as high as 58% for Y^* and 32% for Z^*. These values are relevant for situations when the exact scenario representing disease is not known.

	$q_{0.90}(Y^*)/q_{0.90}(Y)$			$q_{0.90}(Z^*)/q_{0.90}(Z)$		
β	$\phi = 0.3$	$\phi = 0.5$	$\phi = 0.7$	$\phi = 0.3$	$\phi = 0.5$	$\phi = 0.7$
0.70	1.320	1.489	1.587	1.230	1.300	1.325
0.80	1.289	1.453	1.560	1.210	1.300	1.325
0.90	1.257	1.417	1.533	1.196	1.286	1.305
0.95	1.239	1.399	1.505	1.190	1.286	1.305

7 Discussion

The tables demonstrate a substantial effect of length biased sampling on both the mean sojourn time and the mean clinical durations. Three scenarios were considered, corresponding roughly to rather fast disease, moderate disease, and a mixture of fast and rather slow-growing disease (Figures 3, 4, 5), where the correlations between the sojourn times and the clinical durations were 0.29, 0.65, 0.46, respectively, and where the test sensitivity ranged from 0.80 to 0.99 (albeit, with the simplyfing assumption that it is constant over time and for all persons being screened). Interestingly, the effect of the test sensitivity is very minor. The main parameter that seems to drive the increase is the product $\delta\lambda$: so long as the mean sojourn time is *much* larger than the screening interval, the theoretical increase in mean sojourn time due solely to length biased sampling can be kept reasonably small (less than 10% error if the screening interval is no more than 20% of the mean

sojourn time). However, this is likely to be impractical, and even when the screening interval is equal to the mean sojourn time and test sensitivity is high, the mean clinical duration from screen-detected cases can be as high as 7 times that for the general population, due solely to the fact that screening tends to pick up the slower-growing disease and hence collect length-biased cases into the sample. (need more)

References

1. Abramowicz, M.; Stegun, I. (1958), *Handbook of Mathematical Functions*, New York, Dover.

2. Blumenthal S: Proportional sampling in life-length studies, *Technometrics* 1967; **9**(2): 205-218.

3. Chen J, Prorok PC, Graff KM: An age dependent stochastic model of periodic screening: Length bias at a prevalence screen, *Mathematical Biosciences* 1983: **65**, 93-123.

4. Cnaan A: Survival models with two phases and length biased sampling, *Communications in Statistics – Theory and Methods* 1985; **14**(4), 861-886.

5. Cox, DR (1969), Some sampling problems in technology. In: "New Developments in Survey Sampling" (eds. N.L. Johnson and H. Smith Jr.), Wiley, New York, 506-527.

6. Cox DR, Lewis PAW (1972), *The Statistical Analysis of Discrete Time Events*, Oxford Press, London.

7. Fontana RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhm JR, Bernatz PE, Payne WS, Pairolero PC, Bergstalh EJ: Screening for lung cancer: A critique of the Mayo lung project, *Cancer* 1991; 67: 1155–1164.

8. Gohagan JK, Prorok PC, Kramer BS, Cornett JE: Prostate cancer screening in the prostate,

lung, colorectal, and ovarian cancer screening trial of the National Cancer Institute, *Journal of Urology* 1994; 152: 1905–1909.

9. Gupta PL, Tripathi RC: Effect of length-biased sampling on the modeling error, *Communications in Statistics – Theory and Methods* 1990; **19**(4), 1483-1491.

10. Jones MC: Kernel density estimation for length biased data, *Biometrika* 1991; **78**, 511-519.

11. Kafadar K, Prorok PC: A data-analytic approach for estimating lead time and screening benefit based on survival curves in randomized trials, *Statistics in Medicine* 1994; 13, 569-586.

12. Kafadar K, Prorok PC: Computer simulation experiments of randomized screening trials, *Computational Statistics and Data Analysis* 1996; **23**, 263-291.

13. Kafadar K, Prorok PC: Alternative definitions of comparable case groups and estimates of lead time and benefit time in randomized cancer screening trials, *Statistics in Medicine* 2003; 21: 83–111.

14. Kafadar K, Prorok PC: Computational methods in medical decision making: To screen or not to screen? *Statistics in Medicine* 2005; 24: 569–581.

15. Kalbfleisch JD, Lawless JF, Robinson JA: Methods for the analysis and predictioon of warranty claims, *Technometrics* 1991; **33**(3): 273–286.

16. Morrison AS: Sequential pathogenic components of rates, *American Journal of Epidemiology* 1979; **109**(6), 709-718.

17. Nair VN, Wang PL: Maximum likelihood estimation under a successive sampling discovery model, *Technometrics* 1989; **31**(4): 423–436.

18. Scheaffer RL: Size-biased sampling, *Technometrics* 1972; **14**(3), 635-644.

19. Schotz WE, Zelen M: Effect of length sampling bias on labeled mitotic index waves, *Journal of Theoretical Biology* 1971; **32**, 383–404.

20. Shapiro S, Venet W, Strax P, Venet L: *Periodic Screening for Breast Cancer: The Health Insurance Plan Project and its Sequelae, 1963–1986*, Johns Hopkins University Press: Baltimore (1988).

21. Spratt JS, Meyer JS, Spratt JA: Rates of growth of human solid neoplasms, Part I, *Journal of Surgical Oncology* 1995; 60: 137–146.

22. Spratt JS, Meyer JS, Spratt JA: Rates of growth of human solid neoplasms, Part II, *Journal of Surgical Oncology* 1996; 61: 68–83.

23. Vardi Y: Empirical distributions in selection bias models, *Annals of Statistics* 1985; **13**: 178–203.

24. Wang, Mei-Cheng: Hazards regression analysis for length biased data, *Biometrika* 1996; **83**: 343–354.

25. Wang, Mei-Cheng: Length bias, *Encyclopedia of Biostatistics, Volume 3* (eds. Peter Armitage and Theodore Colton), Wiley, New York (1998).

26. Zelen M, Feinleib M: On the theory of screening for chronic diseases, *Biometrika* 1969; **56**, 601–613.

27. Zelen M: Theory of early detection of breast cancer in the general population, *Breast Cancer: Trends in Research and Treatment* (ed. J.C. Heuson, W.H. Mattheiem, M. Rozenweig), Raven Press, New York, 287–301 (1976).

CENTER FOR COMPUTATIONAL MATHEMATICS REPORTS

University of Colorado at Denver and Health Sciences Center
P.O. Box 173364, Campus Box 170
Denver, CO 80217-3364

Fax: (303) 556-8550
Phone: (303) 556-8442
http://www-math.cudenver.edu/ccm

233 Jan Mandel, Lynn S. Bennethum, Jonathan Beezley, Janice L. Coen, Craig C. Douglas, Leopoldo P. Franca, Minjeong Kim, and Anthony Vodacek, "A Wildfire Model with Data Assimilation." June 2006.

234 Jan Mandel and Vaibhav Kulkarni, "Construction of a Level Function for Fireline Data Assimilation." June 2006.

235 Gabriel R. Barrenechea, Leopoldo P. Franca , and Frederic Valentin,"A Petrov-Galerkin Enriched Method: A Mass Conservative Finite Element Method For The Darcy Equation." June 2006.

236 Leopoldo P. Franca, Volker John, Gunar Matthies and Lutz Tobiska, "An Inf-Sup Stable and Residual-Free Bubble Element For The Oseen Equations." June 2006.

237 Jan Mandel, Bedrich Sousedik, and Clark R. Dohrmann, "On Multilevel BDDC." November 2006

238 Janice L. Coen, Jonathan D. Beezley, Lynn S. Bennethum, Craig C. Douglas, Minjeong Kim, Robert Kremens, Jan Mandel, Guan Qin, and Anthony Vodacek, "Wildland Fire Dynamic Data-Driven Application System." November 2006

239 Jan Mandel and Jonathan D. Beezley, "Predictor-Corrector and Morphing Ensemble Filters for the Assimilation of Sparse Data into High-Dimensional Nonlinear Systems." November 2006

240 Jonathan D. Beezley and Jan Mandel, "Morphing Ensemble Kalman Filters." February 2007

241 Jan Mandel, Jonathan D. Beezley, Lynn S. Bennethum, Soham Chakraborty, Janice L. Coen, Craig C. Douglas, Jay Hatcher, Minjeong Kim, and Anthony Vodacek, "A Dynamic Data Driven Wildland Fire Model." Ferbuary 2007

242 Jan Mandel, "A Brief Tutorial on the Ensemble Kalman Filter." February 2007

243 Jonathan D. Beezley and Jan Mandel, "Morphing Ensemble Kalman Filters." May 2007

244 Karen Kafadar, Philip C. Prorok, "Effect of Length Biased Sampling of Unobserved Sojourn Times on the Survival Distribution When Disease is Screen-Detected." June 2007