Hindawi Publishing Corporation
International Journal of Mathematics and Mathematical Sciences
Volume 2011, Article ID 895273, 16 pages
doi:10.1155/2011/895273

Research Article

Risk-Adjusted Control Charts for Health Care Monitoring

Willem Albers

Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Correspondence should be addressed to Willem Albers, w.albers@utwente.nl

Received 29 July 2011; Revised 2 September 2011; Accepted 6 September 2011

Academic Editor: Frank Werner

Attribute data from high-quality processes can be monitored effectively by deciding on whether or not to stop at each time where $r \geq 1$ failures have occurred. The smaller the degree of change in failure rate during out of control one wants to be optimally protected against, the larger the r should be. Under homogeneity, the distribution involved is negative binomial. However, in health care monitoring, (groups of) patients will often belong to different risk categories. In the present paper, we will show how information about category membership can be used to adjust the basic negative binomial charts to the actual risk incurred. Attention is also devoted to comparing such conditional charts to their unconditional counterparts. The latter do take possible heterogeneity into account but refrain from risk-adjustment. Note that in the risk adjusted case several parameters are involved, which will all be typically unknown. Hence, the potentially considerable estimation effects of the new charts will be investigated as well.

1. Introduction

We are interested in processes which exhibit only a (very) small proportion of defectives. Due to ever increasing efforts and standards, such high-quality processes become more and more common in industrial setups. Moreover, for the quite different field of health care monitoring, they are in fact the rule: errors, such as malfunctioning equipment, fatally delayed help, or surgical failures, should occur only (very) rarely. Now review papers on health care monitoring (see, e.g., [1–3] and Sonesson and Bock [4]) strongly suggest to apply SPC methods, in particular control charts, and we shall follow that line here.

The common starting point for monitoring such attribute data is to watch the number of failures in a series of given sampling intervals. However, for high-quality processes, this p-chart may not be the best choice. A first improvement is to switch to a "time-between-events" or "geometric" chart, which uses the number of successes between failures to judge whether

the process has remained in control (IC). See, for example, Liu et al. [5], Yang et al. [6], Xie et al. [7], Ohta et al. [8] and Wu et al. [9], for details. When the process goes out of control (OoC), such geometric charts quickly react to substantial increases of the failure rate p from IC but are admittedly rather slow in detecting moderate changes. Especially in health care applications, this is undesirable, and several of the authors mentioned above have suggested a second improvement step.

Here a decision whether or not to stop is no longer made after each failure but instead only after $r > 1$ failures have occurred. Typically, the smaller the increase of failure rate during OoC one wants to have optimal detection power against, the larger the r should be. This negative binomial chart is analyzed in some detail in Albers [10]. In particular, a simple rule of thumb is presented for selecting r, and the resulting chart is both easy to understand and to apply. However, as subsequently pointed out in Albers [11], a serious complication arises if the underlying homogeneity assumption is not warranted. In industrial applications, it may often—but by no means always—be reasonable to indeed assume one and the same IC failure probability p for each item inspected. But, in medical settings, patients tend to exhibit quite a bit of heterogeneity, and we will regularly have to take such variation between subjects into account.

In Albers [11] the basic situation is considered where in fact all we know is that such heterogeneity does occur. It can, for example, stem from the occurrence of different groups, each with its own IC probability of failure, but we lack further information. The only way, in which it becomes visible, is through an increase of variance, as compared to the homogeneous case. For a discussion of this overdispersion phenomenon, see, for example, Poortema [12] for a general review and Christensen et al. [13] and Fang [14] for earlier applications concerning attribute control charts. In Albers [11] it is demonstrated how the negative binomial charts can be generalized to cover the present overdispersion situation. Essentially the ill-fitting single parameter homogeneous model is widened there into a two-parameter model. In addition to the failure rate p, a second parameter is added, in order to capture the degree of overdispersion. In view of the lack of knowledge about the precise underlying mechanism of the process, this wider family still remains an approximation of reality. But, as demonstrated in Albers [11], the results under overdispersion are far better than those provided by the homogeneous approach.

As was already pointed out in Albers [11], quite a different situation occurs when we do have information about the underlying structure. For example, suppose a number of risk categories can be distinguished, each with its own p_j during IC, and for each incoming patient we register to which class he/she belongs. First of all, such detailed knowledge about the process in principle allows a more accurate analysis. But probably even more important is the fact that it opens the way to applying so-called risk adjustment methods (see, e.g., [15, 16] for an overview, and [17] for a risk-adjusted version of the sets method introduced by Chen [18]). Here the baseline risk of each patient is taken into account in deciding whether the process is still IC. If, for example, a surgeon's performance decreases over time, an OoC signal may nevertheless not be justified if we observe that meanwhile his/her patients are gradually shifting to higher risk categories.

Clearly this is an interesting area, giving rise to quite a few questions, both from a practical and a technical point of view. For example, in practice, one can wonder under what circumstances one should adjust the risk and when one should ignore this possibility and stick to a rigid overall chart. A more technical issue is the following. In risk adjustment several parameters are involved (cf. the p_j above), which typically are unknown and need to be estimated. Now estimation effects for control charts tend to be conveniently ignored in

practice. But if these are studied, they typically turn out to be substantial; for example, Chen et al. [19], mention a 30–90% increase in false alarm rate for a 10% bias in the estimator for p. This estimation topic was studied more systematically in Albers and Kallenberg [20, 21]. There it has been amply demonstrated that the small probabilities involved, such as p, invariably produce large relative errors, and; thus, corrections are called for. In the present situation, we have not one, but several parameters, and hence the effect is likely to be even more serious. But, as Woodall [1] remarks, little work has been done on the effect of the estimation error on the performance of risk-adjusted charts. Consequently, the purpose of the present paper is to remedy this by studying how the negative binomial charts from the simple homogeneous case can be adapted to the situation where risk adjustment is desirable.

As concerns the relation of the methodology proposed here to the existing methods as described in, for example, Steiner et al. [15] and Grigg and Farewell [16, 17], the following remarks are in order. The latter category are of CUSUM-type and as such may be slightly more efficient. In passing note that this actually is a rather subtle matter, as it also seems to depend on the type of performance criterion used (e.g., steady state or not). Nevertheless, using the negative binomial type of approach implies some aggregation of the data over time, which could indeed mean some loss of information compared to the stepwise CUSUM approach. However, precisely this aggregation effect makes the resulting structure less complicated, thus, allowing a detailed analysis of estimation effects, as well as corrections of these. For the CUSUM case (so far), this seems intractable. Hence, the issue in comparing the two types of approach actually is robustness. On the one hand, we have procedures which (maybe) are superior if optimal conditions hold. In the present case that means known parameters, which is almost never realistic. In addition, the impact of just plugging in estimates for these parameters is known to be huge (e.g., causing an average run length during IC which is systematically substantially lower than prescribed). Under such circumstances, it seems an attractive alternative to pay a small insurance premium (in the form of a small loss in detection power) in order to obtain robust procedures which allow control of the validity through adequate corrections for the charts. Incidentally, note that this robustness issue is by no means typical for the application at hand, but is of a rather general nature.

The paper is organized as follows. As far as possible, the technicalities involved are relegated to the appendix section, while the body of the paper provides the main ideas. Section 2 is devoted to introducing briefly the negative binomial chart from Albers [10], which forms our starting point. In Section 3 these charts are subsequently adapted to situations where risk adjustment is called for. The estimation aspect will be the subject of Section 4. For illustrative purposes, throughout the paper examples are presented. Moreover, at the end of the paper, we summarize the application of the proposed chart in a simple set of steps.

2. The Homogeneous Case

Here we briefly introduce the homogeneous case (see [10] for a detailed description). The process being monitored can be characterized through a sequence of independent identically distributed (i.i.d.) random variables (r.v.'s) $D_1, D_2, \ldots$, with $P(D_1 = 1) = 1 - P(D_1 = 0) = p$ in the IC stage. During OoC, p becomes θp for some $\theta > 1$, and we should stop as quickly as possible. As discussed in the Section 1, the option to stop arises each time r failures have been assembled, for some $r \geq 1$. Let X_i, $i = 1, 2, \ldots$, be the successive numbers of Ds involved,

then these X_i are i.i.d. as well, and moreover, distributed as a negative binomial r.v. $X_{r,p} = X$ (indices will be suppressed unless confusion might arise); that is, for $k = r, r+1, \ldots$, we have

$$P(X_{r,p} = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}. \tag{2.1}$$

A stopping signal should occur the first time an $X_i \leq n$, with the lower limit $n = n_{r,p}$ selected such that the false alarm rate (FAR) $F_{r,p}(n) = P(X_{r,p} \leq n)$ equals $r\alpha$, for some small $\alpha > 0$. Then the average run length (ARL) in terms of number of failures during IC equals $r/(r\alpha) = 1/\alpha$ for all r, thus aligning the various negative binomial charts for $r \geq 1$ in a proper way. Consequently, $n = F_{r,p}^{-1}(r\alpha)$, the $r\alpha$th quantile of $F_{r,p}$, which can be obtained numerically.

For the purpose of analyzing the behavior of the lower limit n, a simple and transparent approximation is most useful. This can in its turn also be used for finding an approximation for the ARL during OoC; that is, when p has turned into θp. As this ARL is a function of α, θ, and r, it is in particular interesting to figure out which choice of r is the best for given α and θ. This task has been carried out in Albers [10], to which we refer for a complete description. However, to facilitate independent reading, we do present some of the details in the Appendix section. At this point we just mention that $n \approx \lambda/p$, where λ solves $P(Z_\lambda \geq r) = r\alpha$, in which the r.v. Z_λ is Poisson's with parameter λ. Moreover, by considering a table of numerically obtained optimal values of r (i.e., resulting in the lowest ARL) for the various α and θ of interest and fitting to these $r's$ a simple approximation in terms of α, θ, and $\alpha\theta$, the following easy rule of thumb has been obtained:

$$\tilde{r}^{\text{opt}} = \frac{1}{\alpha(2.6\theta + 2) + 0.01(4\theta - 3)}. \tag{2.2}$$

As can be seen from Table 3 in Albers [10], this simple rule works remarkably well. Quite often the solution from (2.2) is truncated at 5: most of the improvement over the geometric chart (where $r = 1$) has already been achieved at $r = 5$. Moreover, using really large values of r may be considered undesirable in practice anyhow.

As already observed in the Section 1, the underlying parameters typically need to be estimated in practice. In the simple homogeneous case, we merely have to worry about the IC failure rate p. To this end, a Phase I is added, during which the process is observed until m failures have occurred, leading to m geometric $X_{1,p}$'s (cf. (2.1)) and, thus, to the obvious estimator $\hat{p} = 1/\overline{X}$, where $\overline{X} = m^{-1}\Sigma_{i=1}^{m} X_i$. Then we simply replace p by $\hat{p}$ in n (or $\tilde{n}$), after which the actual monitoring can start using the new $\hat{n}$ (or $\hat{\tilde{n}}$). Consequently, application of the chart is almost as easy as that in the known parameter case. However, it does remain to investigate, and possibly to correct, the resulting estimation effects. What happens is that performance characteristics such as FAR (or ARL) are no longer fixed at $r\alpha$ (or $1/\alpha$) but instead will be random in $\overline{X}$. Noting that $U = p\overline{X} - 1$ satisfies $EU = 0$ and $EU^2 = (1-p)/m \approx 1/m$, it follows that the Taylor expansion in terms of U allows adequate appraisal of errors such as $\widehat{FAR} - FAR$. This can be done in terms of bias by taking expectations or in terms of exceedance probabilities by, for example, looking at $P(\widehat{\text{FAR}} > FAR(1+\varepsilon))$. Next, if desired, suitable corrections can be derived. Then slightly lower limits $\hat{n}_c = \hat{n}(1-c)$ are used in which the small positive c is selected such that either the bias is removed or the exceedance

probability will stay below a suitably small upper bound. To illustrate the simple application of the homogeneous chart, we conclude the present section with an explicit example.

Example. Suppose we are monitoring some large population w.r.t. some unpleasant type of event ("a failure"). As long as matters are IC, its incidence rate is $p = 0.001$. Stopping during IC should be rare, so we decide that on average a false alarm once every 200 failures is acceptable, implying that the IC-ARL $= 200$ and $\alpha = 0.005$. Deciding about stopping or continuing at each failure (the geometric chart, with $r = 1$) is known to be not very efficient, so we increase r somewhat, and, for example, take $r = 3$. (Background: the rule of thumb (2.2) tells us that this is optimal at the given α for $\theta = 6$, that is, for an increase during OoC to $p = 0.006$; someone interested in optimality for lower θ should take an even somewhat larger r; for example, $r = 5$ is the best for $\theta = 4$). During IC, each third failure will on average arrive after 3000 observations, and, hence, the chart should signal if this arrival happens much sooner. What "much sooner" should mean under the given conditions has been derived above: the exact lower bound here equals $n = F^{-1}_{3,0.001}(0.015) = 509$. Using λ which solves $P(Z_\lambda \geq 3) = 0.015$ produces $\lambda = 0.509$, and hence $n = \lambda/p = 509$ as well, while the further approximation through (A.2) gives the quite close value 506. Monitoring is now completely straightforward: watch the sequence of realized "third-failure-times" and stop as soon as a number at or below 509 (or 506) is observed. If the IC value of p is in fact unknown, monitoring is preceded by a Phase I sample. First wait till, for example, $m = 100$ failures have been obtained and replace p by the outcome $\widehat{p} = 100/\Sigma_{i=1}^{100} X_i$. If $\widehat{p}$ happens to equal 0.001, the above remains as it was; for any other value, the computation is adapted in a straightforward way.

3. The Heterogeneous Case

Our starting point is the homogeneous case described in Section 2: an underlying sequence of i.i.d. Bernoulli r.v.'s D_i, giving rise to negative binomial r.v.'s for controlling the process. To this situation we now add heterogeneity, and subsequently we investigate how to accommodate this complication, both in the unconditional case, where the underlying information is ignored, and the conditional case, where it is used. Of course, the emphasis will be on the latter situation, where risk adjustment is applied. In fact, the unconditional case has already been dealt with in Albers [11]; we only review it here to provide the proper perspective.

In line with the common notation for bivariate situations, we will use Y's for the main r.v.'s, on which the control decisions are based and X's for the r.v.'s supplying additional information. To be more precise, we will use X as a risk category indicator: $P(X = j) = \pi_j$, $j = 1, \ldots, k$, for some $k \geq 2$. If this is done individually, that is, for each D_i separately, the unconditional situation remains simple. In fact, ignoring or not having the information from the X's means that effectively we still are in the homogeneous case. Each incoming patient simply has the overall failure probability $p = \Sigma \pi_j p_j$, with $p_j = P(D = 1 \mid X = j)$ during IC, and thus the negative binomial charts from Section 2 are still appropriate.

However, matters change if patients arrive in groups, and heterogeneity does lead to nonnegligible overdispersion. In the Appendix section, this is demonstrated in some detail for the relatively simple setting of a fixed number of groups, each of fixed size, to which routinely a standard p-chart would be applied. Moreover, in this context, the difference in behavior between unconditional and conditional charts is analyzed. An excursion to

the situation of continuous rather than attribute data provides additional clarity, as matters for the normal case are somewhat more transparent. The conclusion is that, whenever Y goes OoC, two situations should be distinguished. In the first, the p.d.f. of X (and, thus, the π_j) remains unchanged. Then the two charts react essentially in the same way, with the conditional chart being somewhat more efficient, as it takes the additional information into account. However, once the p.d.f. of X also changes during OoC, the behavior of the charts will diverge. Which of the two provides the right answer depends on whether these changes in X should be ignored or taken into account. (cf. the surgeon from Section 1 who is faced with increasingly more risky patients).

Our present setup is of course even more complicated: group sizes will typically neither be fixed nor equal. As a consequence, the p.d.f. of the waiting times till the rth failure is not only no longer negative binomial but in fact rather intractable in general. For this reason Albers [11] proposes by way of approximation a parametric model containing an additional parameter to cover the overdispersion aspect (c.f. the references mentioned in the Appendix section for accommodating tail length in case of the normal mixture model). In this way an adequate analysis of the unconditional chart indeed becomes possible; see Albers [11] for details.

Having set the background and explained the interpretation, we can now fill in the details for the risk-adjusted chart in a straightforward way. Just as in the homogeneous case, define a new sequence $Y_1, Y_2, \ldots$ on the basis of the D_i's. Here the Y_i are i.i.d. copies of $Y = Y_{r,p}$, the number of D_i to be observed until the rth failure arrives. We decide to stop as soon as a $Y_i \leq n$, for some suitable lower limit $n = n_{r,p}$. In line with Section 2, "suitable" will be interpreted to mean that $P(Y_{r,p} \leq n) = r\alpha$ for some small, given α. To obtain n in the present case, note that while we are waiting for the realization of an Y_i, at each time t, we now have at our disposal the information that $t = \Sigma g_j$, where $g_j = g_{j,t}$ is the number of patients from category j, $j = 1, \ldots, k$. Arguing along the same lines as in Section 2, we obtain (once more see the Appendix section for some of the details) that the simple approximation from the homogeneous case $n = \lambda/p$, with λ solving $P(Z_\lambda \geq r) = r\alpha$, is updated into (cf. (C.2)):

$$n = \Sigma_{j=1}^{k} g_j, \quad \text{with the } g_j = g_{j,n} \text{ such that } \lambda = \Sigma_{j=1}^{k} g_j p_j. \tag{3.1}$$

Note that n from (3.1) can obviously be written as $n = \lambda/(\Sigma\omega_j p_j)$, with weights $\omega_j = g_j/(\Sigma g_j)$. Hence, if we are in the homogeneous case after all; that is, $p_j \equiv p$, we get back $n = \lambda/p$ exactly. Moreover, during IC, the ω_j will tend to be close to the underlying probabilities π_j, and $n = \lambda/p$ will hold approximately for arbitrary p_j as well.

Nevertheless, we should observe that, unlike in the homogeneous case, we no longer have a single fixed n. Each of the realizations Y_i from the above-mentioned sequence $Y_1, Y_2, \ldots$ has its own n_i (and; thus, writing $n = n_{r,p,i}$ would be appropriate now). In fact, we do not even need to obtain all of these n_i explicitly. For example, in those cases where a signal occurs, we have $Y_i \leq n_i$, which means that in this situation at time $t = Y_i$ we still have $\Sigma g_{j,t} p_j \leq \lambda$. Evaluating the actual n_i would imply continuing with the D's corresponding to Y_{i+1}. That would not be the right thing to do; fortunately, it is also superfluous, as the information suffices that the event $\{Y_i \leq n_i\}$ has occurred.

Example (continued). Suppose that for the situation considered in Section 2 additional information, and hence the possibility for risk adjustment, has become available. To keep

matters again as simple as possible, we distinguish just two risk categories: "mild" and "severe." Suppose 10% of the population is severe and their risk is 11 times as high as that of the mild cases. Hence, $k = 2, \pi_2 = 0.1, p_2 = 11p_1$, and thus $\Sigma\pi_j p_j = 2p_1 = p$, resulting in $p_1 = 0.0005$ and $p_2 = 0.0055$. After (3.1) it was observed that the risk-adjusted chart replaces the homogeneous choice $n = \lambda/p$ by $n = \lambda/(\Sigma\omega_j p_j)$, in this way, taking the actually observed weights into account. Here this means using $n = (\lambda/p)\{2/(1 + 10\omega_2)\}$, which boils down to $509\{2/(1 + 10\omega_2)\}$. During IC, ω_2 will be close to $\pi_2 = 0.1$, and hence n will be close to 509, as in the homogeneous case. Nevertheless, monitoring becomes slightly less trivial than before: now not only the sequence of realized third failure times but also the corresponding fractions of severe cases ω_2 need to be recorded. For each data pair (Y_i, ω_{2i}), it is checked whether $Y_i \leq 509\{2/(1 + 10\omega_{2i})\}$. For example, an outcome $(498, 0.15)$ produces a signal in the homogeneous case, ("$498 \leq 509$"), but not in the risk-adjusted situation ("$498 \geq 472 = 509/\{2/(1 + 1.5)\}$"). In the latter case the occurrence of the third failure at a very early moment is deemed acceptable after all in view of the somewhat larger than expected presence of severe cases.

The step in (3.1) is essentially all that is needed to deal with the risk adjusted version of the chart. In the Appendix section, it is demonstrated how the approximation steps for, for example, λ and the ARL carry over. Of course, if the process goes uniformly OoC, in the sense that all p_j are replaced by θp_j, matters are most straightforward. But even if each p_j has its own θ_j, it remains easy to adapt the previously obtained expressions (cf. (C.3)). It is also explicitly demonstrated how the above-mentioned weights $\omega_j = g_j/(\Sigma g_j)$ influence matters. If these remain close to the π_j, the risk-adjusted chart behaves in the same manner as its unconditional counterpart. It only is slightly more efficient. However, if the ω_j's are quite different, the behaviour will diverge. To illustrate this, just as in the normal example from the fixed case, an explicit example is given in the Appendix section of an OoC situation for Y which is completely due to the change in X. The risk-adjusted chart then indeed continues to consider the process as being IC. To illustrate matters, we conclude the present section with α numerical example.

Example (continued). In the setup considered before, let us next consider the OoC situation. First let $\theta = 2$ uniformly; that is, the mild category gets $\theta p_1 = 0.001$ and the severe one $\theta p_2 = 0.011$. From, for example, Albers [11], we obtain that the homogeneous chart has an ARL of 36 for this case, and, according to the above, this will continue to hold here as well. Next consider a nonuniform example: let $\theta_1 = 7/9$, and let $\theta_2 = 3$, then $\Sigma\pi_j\theta_j p_j = 2p$, and; hence, once more $\theta = 2$. If during OoC the p.d.f. of X remains the same, the π_i's do not change, and ω_2 will still be close to $\pi_2 = 0.1$. Then the risk-adjusted chart will continue to behave like the homogeneous one, that is, with an ARL of about 36 at this $\theta = 2$. However, if instead the p_j's remain the same and the θ_j's are used to transform $(\pi_1, \pi_2) = (0.9, 0.1)$ into $(\theta_1\pi_1, \theta_2\pi_2) = (0.7, 0.3)$, the chart will shift towards using a lower limit n close to $\lambda/(2p) = 509/2$. Moreover, as $\theta^* = 1$ in (C.3) for this particular choice, its ARL will remain at about the IC value 200. Clearly, from a risk adjustment perspective, this is precisely what should happen. The mild patients still have $p_1 = 0.0005$, and the severe ones still have $p_2 = 0.0055$. What has changed is that the latter category has shifted from 10% to 30% of the total. Hence, not the quality of the performance has deteriorated, but rather that of the incoming patients (cf. the surgeon example in Section 1).

4. Estimation Effects

Typically the underlying parameters of control charts are not known in practice. Here this means that we will have to estimate not just the overall $p = \Sigma\pi_j p_j$, but the individual p_j as well. As concerns p, in Albers [10] a Phase I sample of "size" m was used for this purpose, meaning that we observe $D_1, D_2, \ldots$ until m failures have arrived. Note that m, and hence the length $Y_{m,p}$ of this sample as well, is independent of r, implying that the comparison between charts for different r remains fair also w.r.t. estimation. Next, p was simply estimated by $\widehat{p} = m/Y_{m,p}$. In the present context we can use this same sample, but in a more detailed way, as follows. Let H_j be its number of patients from category j (i.e., $\Sigma_{j=1}^{k} H_j = Y_{m,p}$), and in addition let D_{ji}, $i = 1, \ldots, H_j$ denote the corresponding D's. Then we have as a straightforward choice for estimating the p_j:

$$\widehat{p}_j = \frac{\Sigma_{i=1}^{Hj} D_{ji}}{H_j}, \quad j = 1, \ldots, k. \tag{4.1}$$

Of course, formally there is a problem in (4.1), as each H_j can be 0 with positive probability. Using a slightly modified definition could remedy this. However, we shall not bother to do so, as the probabilities involved are exponentially small. Moreover, if too small h_j's are observed, this anyhow indicates that the design may not be right, and additional effort is required before monitoring can begin. Given $H_j = h_j$, we have $E\widehat{p}_j = p_j$ and $\mathrm{var}(\widehat{p}_j) = p_j(1 - p_j)/h_j$. As $EH_j = \pi_j EY_{m,p} = m\pi_j/p$, it follows that, ignoring terms involving p_j^2,

$$\frac{\widehat{p}_j}{p_j} - 1 \approx \mathrm{AN}\left(0, \frac{p}{m\pi_j p_j}\right), \tag{4.2}$$

with "AN" denoting asymptotic normality. Hence, if the Phase I sample is chosen in this way, the estimators $\widehat{p}_j$ are indeed well behaved, in the usual sense of having a relative error which is $O_P(m^{-1/2})$. Only if the contribution $\pi_j p_j$ of a certain category j is really small compared to $p = \Sigma\pi_i p_i$, the coefficient involved will be large.

In fact, the above is all that is needed to transform the chart into its estimated version: just replace in (3.1) the p_j by their estimated counterparts $\widehat{p}_j$ from (4.1). To be precise, we use a lower limit $\widehat{n}$ defined by

$$\widehat{n} = \Sigma_{j=1}^{k} \widehat{g}_j, \quad \text{with the } \widehat{g}_j = \widehat{g}_{j,n} \text{ such that } \lambda = \Sigma_{j=1}^{k} \widehat{g}_j \widehat{p}_j, \tag{4.3}$$

where once again λ is such that $P(Z_\lambda \geq r) = r\alpha$. The alternative notation then becomes $\widehat{n} = \lambda/(\Sigma\widehat{\omega}_j\widehat{p}_j)$, with weights $\widehat{\omega}_j = \widehat{g}_j/(\Sigma\widehat{g}_j)$. Once this lower limit $\widehat{n}$ has been obtained, the actual monitoring can start: each time wait till the rth failure, and if this occurs at or before $\widehat{n}$, a signal results. Hence, straightforward application of the estimated chart remains easy.

To investigate the effects of the estimation step, we proceed as follows. As remarked in Section 2, FAR will no longer be fixed at some given value $r\alpha$ nor will ARL be precisely equal $1/\alpha$. These performance characteristics have now become the random variables $\widehat{\mathrm{FAR}}$

and $\widehat{\text{ARL}}$, respectively. They depend on the estimators $\widehat{p}_j$ and consequently fluctuate around the intended values. A rather detailed analysis of the consequences for the homogeneous case can be found in Albers [10]. To avoid repetition, we shall be much more brief here and mainly focus on the additional complication caused by the fact that the estimation step is now split into k categories. See once more the Appendix section for the details. It is evaluated how large the size m of the Phase I sample should be in order to ensure that the exceedance probabilities discussed at the end of Section 2 will fall below a prescribed small quantity δ. If the resulting m is too large for use in practice, a small correction c is calculated, such that for a given m the desired δ can be met after all by using the slightly more strict $\widehat{n}_c = \widehat{n}(1-c)$ rather than $\widehat{n}$ itself. A complication here in comparison to the homogeneous case is the presence of an additional quantity $\tau \geq 1$ (see (D.5)), which represents the unbalance caused by (possible) differences between the ideal π_j and the actually occurring ω_j.

To illustrate matters we again present an explicit example.

Example (continued). Once again, we add a complication to our ongoing example: now the values of p_1 and p_2 are no longer known. Suppose, therefore, that we first wait till $m = 100$ failures have occurred and then use the resulting Phase I sample to estimate these p_j (see (4.1)). Just as in the example of Section 2, the application of the chart remains straightforward: simply plug in the $\widehat{p}_j$ to replace the unknown p_j. However, if we also want to study the impact of the estimation step, and possibly correct for it, a bit more effort is needed (we shall refer to the Appendix section for the required formulae). The point is that, in the subsequent monitoring phase, we may still aim at an ARL of 200, but we have to accept that the actual $\widehat{\text{ARL}}$ may differ. In particular, this can substantially be smaller than 200, leading to a lot more frequent false alarms than anticipated. To monitor not only the process but this effect as well, we can, for example, look at the probability of a deviation of more than 20%, that is, of the $\widehat{\text{ARL}}$ falling below 160. In terms of (D.4), this is precisely P_{Exc} with $1/(1+\varepsilon) = 0.8$, and; thus, $\varepsilon = 0.25$. As moreover $r = 3$, it then follows from (D.6) that $P_{\text{Exc}} \approx 1 - \Phi(2.5/(3\gamma\tau))$. Since γ is close to, but smaller than, 1, a close upper bound for P_{Exc} is $1 - \Phi(2.5/(3\tau))$. As long as X is IC (regardless of whether the same holds for Y or not), τ will be close to 1 as well, in which case this upper bound approximately boils down to $1 - \Phi(2.5/3) = 0.20$. If such a 20% probability for a more than 20% too low ARL is acceptable, we can continue to the monitoring phase. Otherwise m should be larger than 100; to be precise, $m \geq (12u_\delta)^2$ is needed to get $P_{\text{Exc}} \leq \delta$. Or, alternatively, for m fixed at 100, this can be achieved by using $\widehat{n}_c$ with $c = (100)^{1/2}u_\delta - 0.25/3$ (cf. (D.7)). For, for example, $\delta = 0.10$, we have $u_\delta = 1.28$ and $m \geq 236$ and $c = 0.045$ result. If in fact X is OoC, τ will be larger than 1, and we need to use $m \geq (12\tau u_\delta)^2$ instead. For example, consider once more the situation from the previous example, where $\pi_1 = 0.9,\ p_1 = p/2, \pi_2 = 0.1,\ p_2 = 11p/2$, and subsequently $\omega_1 = 0.7,\ \omega_2 = 0.3$. Then in addition to $\Sigma\pi_j p_j = p,\ \Sigma\omega_j p_j = 2p$, we obtain $\Sigma[\omega_j^2/\pi_j]p_j = 5.22p$, which leads through (D.5) to $\tau^2 = 1.31$. Hence, to still get $\delta = 0.20$, m should now be 131, rather than just 100. This increase is still rather mild, but it is easy to see that higher increases of m can be necessary. For example, slightly generalize our example into $\pi_1 = 1-q,\ \pi_2 = q$ for some small $q > 0$, while otherwise keeping $p_2/p_1 = 11$ and $(\omega_1, \omega_2) = (0.7, 0.3)$. In that case $p_1 = p/(1+10q)$, producing $\Sigma\pi_j p_j = p, \Sigma\omega_j p_j = 4p/(1+10q)$ and $\Sigma[\omega_j^2/\pi_j]p_j = \{0.49/(1-q) + 0.99/q\}p/(1+10q)$. This leads to $\tau^2 \approx (1+10.5q)/(16q)$, which, for example, for $q = 0.05$ equals 1.9 and for $q = 0.02$ already 3.8. Of course, the latter case is rather extreme, as the frequency of severe cases has increased by a factor 15.

(1) Before the monitoring phase starts, take the following preliminary steps:
  (a) Select a desired ARL = $1/\alpha$ and a degree of change θ during OoC that should be optimally protected against.
  (b) Apply rule of thumb (2.2) to obtain r (typically truncate at 5 in practice).
  (c) Find λ such that $P(Z_\lambda \geq r) = r\alpha$, where Z_λ is Poission, or simply use its approximation $\tilde{\lambda}$ from (A.2).
  (d) Wait till m failures have occurred. Take for example, $m = 100$ (or use Section 4 (e.g., see (D.6)) to make a more elaborate choice).
  (e) From this Phase I sample, evaluate the fraction of failures $\hat{p}_j$ for each of the categories $j = 1, \ldots, k$.
(2) Now wait till Y_1, the moment at which the rth failure occurs.
(3) Obtain the corresponding numbers g_j from category j (i.e., $\Sigma_{j=1}^{k} g_j = Y_1$).
(4) Give a signal if $\Sigma_{j=1}^{k} g_j \hat{p}_j \leq \lambda$; otherwise go back to Step 2, leading to $Y_2, Y_3, \ldots$.

Algorithm 1

To conclude this section, for convenience, we briefly summarize the steps involved in applying the new chart (see Algorithm 1).

Appendices

A. Approximations for the Negative Binomial Chart

In addition to a numerical solution for n, it is desirable to derive a simple approximation as well, for example, to make transparent how the function $n_{r,p,\alpha}$ behaves. Now

$$F_{r,p}(n) = P(X_{r,p} \leq n) = P(Y_{n,p} \geq r) \approx P(Z_{np} \geq r), \tag{A.1}$$

with here $Y_{n,p}$ binomial with parameters n and p and Z_{np} Poisson with parameter $\lambda = np$. Hence, $n \approx \lambda/p$, with λ solving $P(Z_\lambda \geq r) = r\alpha$. For $p \leq 0.01$, $r \leq 5$, and $\alpha \leq 0.01$ (which region amply suffices for our purposes), it is demonstrated in Albers [10] that this λ can be approximated by

$$\tilde{\lambda} = \alpha_r(1 + \zeta_r), \quad \text{with } \zeta_r = \frac{\alpha_r}{r+1} + \frac{1}{2}\alpha_r^2 \frac{3r+5}{(r+1)^2(r+2)}, \tag{A.2}$$

in which $\alpha_r = (r!r\alpha)^{1/r}$. The resulting approximation $\tilde{n} = \tilde{\lambda}/p$ turns out to work well over the region considered.

During OoC we have ARL = $\text{ARL}_{r,\theta} = r/F_{r,\theta p}(n_{r,p}) \approx r/P(Z_{\theta\lambda} \geq r)$, with still λ such that $P(Z_\lambda \geq r) = r\alpha$. Again an approximation is feasible:

$$\tilde{\text{ARL}} = \frac{r}{1 - \exp(-\theta\alpha_r)\left(1 + \theta\alpha_r + \cdots + \left((\theta\alpha_r)^{r-2}/(r-2)!\right) + (\theta\alpha_r)^{r-1}((1 - \theta\alpha_r\zeta_r)/(r-1)!)\right)}, \tag{A.3}$$

with α_r and ζ_r as in (A.2). It is adequate for the (p, r, α)-region above and $3/2 \leq \theta \leq 4$. The improvement achieved by increasing r beyond 1 can be nicely judged by looking at $h_r = h_{r,\theta} = \mathrm{ARL}_{1,\theta}/\mathrm{ARL}_{r,\theta}$. Due to the alignment during IC, these functions all start with value 1 at $\theta = 1$, after which they increase considerably. Only for really large θ, the decrease to the limiting value $1/r$ sets in.

B. Illustration for a Fixed Number of Groups of Fixed Size

Let us assume in this subsection that the patients from a given risk category arrive in groups of fixed size t, for some $t \geq 1$. (Hence, $t = 1$ stands for the special case of individual arrivals). With the risk category indicator X satisfying $P(X = j) = \pi_j, j = 1, \ldots, k$, the number of defectives Y in such a group hence satisfies $Y \mid X = j \, \mathrm{bin}(t, p_j)$. Suppose we wait till h such groups have arrived, and thus $n = ht$ patients have been seen. Since $E(Y \mid X) = tp_X$ and $\mathrm{var}(Y \mid X) = tp_X(1 - p_X)$, it immediately follows that $EY = tp = t\Sigma\pi_j p_j$ and $\sigma_Y^2 = t(p - Ep_X^2) + t^2\mathrm{var}(p_X) = tp(1 - p) + t(t - 1)\mathrm{var}(p_X)$. For controlling the process in this simple setup, we might want to apply the standard p-chart, based on $S = \Sigma_{i=1}^n Y_i$. But, since $\sigma_S^2 = h\sigma_Y^2 = n\{p(1 - p) + (t - 1)\mathrm{var}(p_X)\}$, it is immediate that an overdispersion effect becomes apparent as soon as $t > 1$, that is, in the case of "real" groups. Now, $\mathrm{var}(p_X) = \Sigma\pi_j(p_j - p)^2$ and we are focusing on high-quality processes, meaning that p and the p_j are (very) small. Consequently, the term $\mathrm{var}(p_X)$, being of order p^2, might still seem negligible. However, if t is of order comparable to n, as will often be the case, the first term in σ_S^2 behaves as np and the overdispersion correction as $(np)^2$. Typically, such terms will be of the same order of magnitude, and heterogeneity does have an impact.

Hence, even in the unconditional case (where the underlying information is unavailable or simply ignored), a suitable modification should be applied to the standard p-chart. To be more precise, the usual upper limit it uses is $np + u_\alpha\{np(1 - p)\}^{1/2}$, where $u_\alpha = \Phi^{-1}(1 - \alpha)$, with Φ denoting the standard normal p.d.f. Clearly, this should now be replaced by $np + u_\alpha\{n[p(1 - p) + (t - 1)\mathrm{var}(p_X)]\}^{1/2}$. Next, to make the step towards the conditional case (i.e., the risk adjusted counterpart), note that we can also write

$$S = \Sigma_{j=1}^k \Sigma_{i=1}^{G_j} Y_{ji}, \tag{B.1}$$

with $(G_1, \ldots, G_k)$ multinomial $(h, \pi_1, \ldots, \pi_k)$ and $Y_{ji} \, \mathrm{bin}(t, p_j)$. Performing control conditional on the observed x_i implies that in the risk-adjusted case S is compared to the upper limit $t\Sigma g_j p_j + u_\alpha\{t\Sigma g_j p_j(1 - pj)\}^{1/2}$. To appreciate the difference in behavior between the unconditional and the conditional version of the chart, just drop the conditioning on the G_j and use that $EG_j = h\pi_j, \mathrm{var}(G_j) = h\pi_j(1 - \pi_j)$ and $\mathrm{cov}(G_i, G_j) = -h\pi_i\pi_j$. It then readily follows that the expectation $n\Sigma\pi_j p_j(1 - p_j)$ of the conditional variance is, thus, increased by the nonnegligible amount $\Sigma(tp_j)^2 h\pi_j(1 - \pi_j) - \Sigma\Sigma_{i \neq j} t^2 p_i p_j h\pi_i\pi_j = nt \, \mathrm{var}p_X$. This result (as it obviously should) agrees with the form of σ_S^2 derived above for the unconditional case.

Consequently, the picture is as follows: as long as the π_j's remain unchanged once Y goes OoC, both charts react similarly, with the risk-adjusted version being the more precise one (as it has a smaller variance). However, if going OoC affects not only Y but also X (and, thus, the π_j), the behavior will diverge, and the choice will become dependent on the aim one has in mind.

To illustrate this conclusion more clearly, we conclude this subsection by making a two-step excursion to the corresponding normal case of monitoring the mean of a continuous process variable by means of a Shewhart chart. Hence, Y now is $N(\mu, \sigma^2)$, that is, has p.d.f. $\Phi((x-\mu)/\sigma)$, and an upper limit $\mu + u_\alpha\sigma$ is used in the homogeneous case. In the presence of the group indicator X, this setup transforms into $Y \mid X = j$ being $N(\mu_j, \sigma_j^2)$. For the unconditional case we then readily obtain that $\mu = \Sigma\pi_j\mu_j$ and $\sigma^2 = \Sigma\pi_j\{\sigma_j^2 + (\mu_j - \mu)^2\}$, but we do note that Y now has a mixture distribution and, hence, is no longer normal. Such violations of the normality assumption in fact occur quite often in practice, and Albers et al. [22, 23] demonstrated how using an additional parameter concerning the tail length of the actual distribution can adequately remedy the ensuing model error. On the other hand, for the conditional case matters remain pretty trivial, just use $\mu_j + u_\alpha\sigma_j$ for the appropriate j as upper limit. Comparison of the two versions runs completely parallel to that for the attribute data above.

To obtain the intended additional clarity, a second step is needed: instead of using a category indicator, now let X be normal as well. Hence, we have pairs (Y, X) which are bivariate normal $N(\mu_Y, \mu_X, \sigma_Y^2, \sigma_X^2, \rho)$. Unconditional monitoring then means using the upper limit $\mu_Y + u_\alpha\sigma_Y$, while working given the observed outcomes x means applying $\mu_Y + \rho\sigma_Y(x - \mu_X)/\sigma_X + u_\alpha(1-\rho^2)^{1/2}\sigma_Y$, in view of the fact that $Y \mid X = x$ is $N(\mu_Y + \rho\sigma_Y(x-\mu_X)/\sigma_X, (1-\rho^2)\sigma_Y^2)$. Such a method of using auxiliary information has recently been discussed by Riaz [24]. The advantage of this bivariate case is that it allows a very simple comparison between the two approaches. For let OoC now imply that the (Y_i, X_i) will be $N(\mu_Y + d_Y\sigma_Y, \mu_X + d_X\sigma_X, \sigma_Y^2, \sigma_X^2, \rho)$, then the conditional approach has an ARL $= 1/\{1 - \Phi(u_\alpha - \tilde{d}_Y)\}$, where

$$\tilde{d}_Y = \left(1 - \rho^2\right)^{-1/2}(d_Y - \rho d_X). \tag{B.2}$$

Indeed, if $d_X = 0$ that is, the X_i's remain unchanged whenever the Y_i's go OoC, then clearly $\tilde{d}_Y > d_Y$, unless $\rho = 0$. Hence, in this situation, inspection using the auxiliary information from the X_i leads to lower ARL and, thus, to better performance, as argued in Riaz [24]. However, if d_X is positive as well, $\tilde{d}_Y$ can be smaller than d_Y and even 0 (just let $d_X = d_Y/\rho$), meaning that no OoC situation is perceived. Whether this is right or wrong just depends on the perspective used. If $d_Y = \rho d_X$, the behavior of Y is shifted "only" because of the shift in the underlying X. In that sense "nothing" has happened, and not reacting may well be the appropriate response. However, if the behavior of Y should be judged against a fixed standard, irrespective of possible shifts in X, one should clearly stick to the unconditional version by using $\mu_Y + u_\alpha\sigma_Y$.

C. Approximations for the Heterogeneous Case

If we denote a binomial r.v. with parameters s and q by $\tilde{Y}_{s,q}$, then, in analogy to (A.1), we have, with $g_j = g_{j,n}$,

$$P(Y_{r,p} \le n) = P\left(\Sigma_{j=1}^{k}\tilde{Y}_{g_j,p_j} \ge r\right) = r\alpha. \tag{C.1}$$

For $r > 1$, typically n will be large, and, if k is (relatively) small, the g_j will be large as well. Hence, the Poisson approximation step from (A.1) can be used here again. In fact, quite

conveniently it remains possible to keep using λ such that $P(Z_\lambda \geq r) = r\alpha$; only the relation between the lower limit n and this λ becomes slightly more intricate than in the homogeneous case:

$$n = \Sigma_{j=1}^{k} g_j, \quad \text{with the } g_j = g_{j,n} \text{ such that } \lambda = \Sigma_{j=1}^{k} g_j p_j. \tag{C.2}$$

(More formally: look for the largest value of n such that the corresponding $g_j = g_{j,n}$ satisfies $\Sigma_{j=1}^{k} g_j p_j \leq \lambda$. However, as the p_j's are small, this distinction is rather futile.)

Till now we excluded the case $r = 1$, as n is not necessarily large there and the Poisson step might not be warranted. Just as in the homogeneous situation, this boundary case can be simply solved exactly. In fact, $P(Y_{1,p} \leq n) = P(\Sigma_{j=1}^{k} \tilde{Y}_{g_j,p_j} \geq 1) = 1 - P(\tilde{Y}_{g_j,p_j} = 0 \ \forall \ j) = 1 - \Pi(1 - p_j)^{g_j} = \alpha$. This leads to $\Sigma g_j \log(1 - p_j) = \log(1 - \alpha)$ and, thus, to, for example, $\Sigma g_j p_j \approx -\log(1 - \alpha)$, which is in line with the result from (C.2) (see [10] for details).

Hence, unlike in the unconditional case, where the Poisson approximation has to be replaced by a negative binomial one, most of the results during IC from the homogeneous case carry over to the risk-adjusted version and the modification needed is actually already covered by (C.2). For example, all results from Albers [10] about approximating λ (cf. (A.2)) can be used directly. Once the process goes OoC, in principle this carrying over continues, certainly if we translate the step from p to θp in a uniform way into replacing each p_j by θp_j. Then ARL = $\text{ARL}_{r,\theta} = r/P(Y_{r,\theta p} \leq n_{r,p})$, in which $P(Y_{r,\theta p} \leq n_{r,p}) = P(\Sigma_{j=1}^{k} \tilde{Y}_{g_j,\theta p_j} \geq r) \approx P(Z_{\theta\lambda} \geq r)$, with still λ such that $P(Z_\lambda \geq r) = r\alpha$. Hence, an approximation like $\widetilde{\text{ARL}}$ from (A.3) continues to hold as well.

More generally, going OoC will mean that p_j becomes $\theta_j p_j$, with the θ_j such that still $\theta = \Sigma \pi_j \theta_j p_j / \Sigma \pi_j p_j > 1$. Using (C.1) again, we obtain that ARL $\approx r/P(Z_{\theta^*\lambda} \geq r)$, with once more λ such that $P(Z_\lambda \geq r) = r\alpha$, but now

$$\theta^* = \frac{\Sigma_{j=1}^{k} \omega_j \theta_j p_j}{\Sigma_{j=1}^{k} \omega_j p_j}. \tag{C.3}$$

Note that if the p.d.f. of the X_i remains unchanged when the D_i's go OoC, the weights ω_j will remain close to the π_j, and; thus, θ^* from (C.3) will in fact be close to θ. Hence, results from the homogeneous case (such as (A.3)) continue to hold approximately. In other words, the risk-adjusted chart indeed shows behavior quite similar to that of the unconditional chart for the individual case. Moreover, it is somewhat more precise than the unconditional chart from the group arrival case, the difference being that the latter is based on a negative binomial rather than a Poisson p.d.f.

To demonstrate that matters can become quite different when the p.d.f. of X is affected as well during OoC, we argue as follows. First of all, note that there is no need that $\theta_j > 1$ for all j. To see this, without loss of generality, suppose that the p_j's are ordered into increasing order of magnitude. Now assume that we have an increasing sequence of θ_j such that not only the π_j but also the $\pi_j \theta_j$ are in fact probabilities. For such θ_j indeed $\theta = \Sigma \pi_j \theta_j \text{p}_j / \Sigma \pi_j p_j > 1$ will hold, as follows, for example, by noting that monotone likelihood ratio implies increasing expectation [25, page 74]). Hence, this choice can serve as yet another example of the case discussed in the previous paragraph: the X_i's remain unchanged, the p_j's turn into $\theta_j p_j$, leading to an overall OoC factor θ.

However, we can also choose to associate the θ_j in $\pi_j\theta_j p_j$ with π_j rather than with p_j. In other words, the weights ω_j are shifted from π_j to $\pi_j\theta_j$, whereas the p_j remain as they are. Consequently, the conditional chart uses $\omega_j = \pi_j\theta_j$ and $\theta_j p_j = p_j$ in (4.3) and, thus, arrives at $\theta^* = 1$. Hence, after risk-adjustment, everything is still perfectly IC, and the chart sees no reason for action at all. On the other hand, the unconditional charts keeps observing that $\theta = \Sigma\pi_j\theta_j p_j / \Sigma\pi_j p_j > 1$ and will tend to react. As argued before, both answers are right; only the underlying questions differ.

D. Effects of and Corrections for the Estimation Step

In analogy to (C.2), we obtain for the present situation that

$$\widehat{\mathrm{FAR}} = P(Y_{r,p} \leq \widehat{n}) = P\left(\Sigma_{j=1}^{k} \tilde{Y}_{\widehat{g}_j, p_j} \geq r\right) \approx P(Z_{\widehat{\lambda}} \geq r), \tag{D.1}$$

with $\widehat{\lambda} = \Sigma\widehat{g}_j p_j$. From (4.3) it follows that $\widehat{\lambda} = \lambda + \Sigma\widehat{g}_j(p_j - \widehat{p}_j)$ and, thus, that $\widehat{\lambda} = \lambda(1 + U)$, where

$$U = \Sigma_{j=1}^{k} \frac{\widehat{g}_j \widehat{p}_j}{\Sigma_{j=1}^{k} \widehat{g}_j \widehat{p}_j} \left(\frac{p_j}{\widehat{p}_j} - 1 \right). \tag{D.2}$$

This is precisely the same structure as we already had in (4.4) from Albers [10], the only difference being that there U simply equals $p/\widehat{p} - 1$. Hence, the subsequent steps are quite parallel, and we shall not go into much detail. The idea again is to expand $\widehat{\mathrm{FAR}} - \mathrm{FAR} \approx P(Z_{\widehat{\lambda}} \geq r) - P(Z_\lambda \geq r)$ w.r.t. U. Since $dP(Z_\lambda \geq r)/d\lambda = P(Z_\lambda = r - 1) = r(Z_\lambda = r)/\lambda$, we arrive at

$$\widehat{\mathrm{FAR}} - \mathrm{FAR} \approx rUP(Z_\lambda = r) = (\gamma r U)\mathrm{FAR}, \tag{D.3}$$

where $\gamma = P(Z_\lambda = r)/P(Z_\lambda \geq r)$. According to Lemma 4.1 from Albers [10], this γ satisfies $1 - \lambda/(r+1) < \gamma < 1$, implying that it will typically be close to 1.

From (D.3) we can readily evaluate desired quantities of interest for judging the impact of the estimation step, such as the exceedance probability:

$$P_{\mathrm{Exc}} = P\left(\frac{\widehat{\mathrm{FAR}}}{\mathrm{FAR}} - 1 > \varepsilon \right) = P\left(\widehat{\mathrm{FAR}} > r\alpha(1 + \varepsilon) \right), \tag{D.4}$$

in which ε is some small positive constant, like 0.25. Note that, since $\widehat{\mathrm{ARL}} = r/\widehat{\mathrm{FAR}}$, we can also write $P_{\mathrm{Exc}} = P(\widehat{\mathrm{ARL}} < (1/\alpha)/(1 + \varepsilon))$, so both types of performance characteristics are dealt with simultaneously through (D.4). Indeed, from (D.3), it immediately follows that $P_{\mathrm{Exc}} \approx P(U > \varepsilon\alpha / P(Z_\lambda = r)) = P(U > \varepsilon/(\gamma r))$. In the homogeneous case, we simply have that $U = p/\widehat{p} - 1$ is $AN(0, m^{-1})$, and thus the corresponding $P_{\mathrm{Exc}} = 1 - \Phi(m^{1/2}\, \varepsilon/(\gamma r))$ (see (4.10) in Lemma 4.3 of [10]). For the present case, we combine (4.2) and (D.2). First note that $U = \Sigma_{j=1}^{k} \widehat{\omega}_j \widehat{p}_j (p_j/\widehat{p}_j - 1) / \Sigma_{j=1}^{k} \widehat{\omega}_j \widehat{p}_j \approx \Sigma_{j=1}^{k} \omega_j p_j (p_j/\widehat{p}_j - 1) / \Sigma_{j=1}^{k} \omega_j p_j$ and also that (4.2) holds

not only for $\widehat{p}_j/p_j - 1$ but also for $p_j/\widehat{p}_j - 1$. Consequently, once more ignoring quadratic terms in the p_j, we obtain that the present U is $AN(0, m^{-1}\tau^2)$, where

$$\tau^2 = \left(\Sigma_{j=1}^{k} \frac{\omega_j^2}{\pi_j} p_j\right) \frac{\Sigma_{j=1}^{k} \pi_j p_j}{\left(\Sigma_{j=1}^{k} \omega_j p_j\right)^2}. \tag{D.5}$$

Hence, it follows that

$$P_{\text{Exc}} \approx P\left(U > \frac{\varepsilon}{\gamma r}\right) \approx 1 - \Phi\left(\frac{m^{1/2}\varepsilon}{\gamma r \tau}\right), \tag{D.6}$$

with τ as in (D.5).

It is immediate from (D.6) that ensuring $P_{\text{Exc}} \leq \delta$ for a certain small δ, like 0.10 or 0.20, requires $m \geq (\gamma r \tau u_\delta)^2/\varepsilon^2$. Moreover, by for example applying Cauchy-Schwarz to the r.v.'s $V_1 = \omega_X P_X^{1/2}/\pi_X$ and $V_2 = p_X^{1/2}$, it is clear that $\tau \geq 1$, and; thus, $P_{\text{Exc}} \geq P_{\text{Exc,Hom}}$, with equality occurring only if $\omega_j = \pi_j$. As we observed in Section 3, the latter situation will be approximately true during IC and, moreover, during OoC, when only Y is affected and X is not. However, as soon as the behaviors of the risk-adjusted and the unconditional charts start to differ, this will mean that the ω_j's are no longer close to the π_j, and then the exceedance probability will start to be larger than its homogeneous counterpart. The verbal explanation is quite straightforward: as remarked after (4.2), the relative precision of the $\widehat{p}_j$ increases in $\pi_j p_j$. As long as the g_j's are such that the weights ω_j's are close to the supposed π_j, this effect is adequately balanced in (D.2). However, once rare categories become more prominent, this balance is disturbed.

Through (D.6) we can check the behavior of the estimated chart and prescribe the minimal m to ensure a desired bound δ on P_{Exc}. However, if this m cannot be realized and we are stuck with some given lower value, this bound can also be achieved by using a somewhat more strict lower limit $\widehat{n}_c = \widehat{n}(1-c)$, with $\widehat{n}$ as in (4.3) and $c > 0$ small. Then $\widehat{\lambda} = \lambda(1+U)$ from (D.1) becomes $\widehat{\lambda}_c = \lambda(1+U)(1-c)$, and; thus, U is replaced by $U - c$ in what follows, leading through (D.6) to

$$P_{\text{Exc}} \approx P\left(U > c + \frac{\varepsilon}{\gamma r}\right) \approx 1 - \Phi\left(m^{1/2}\left(c + \frac{\varepsilon}{\gamma r \tau}\right)\right). \tag{D.7}$$

A requirement like $P_{\text{Exc}} = \delta$ can now be enforced by choosing $c = m^{-1/2}u_\delta - \varepsilon/(\gamma r \tau)$ (cf. (4.11) from Albers [10]). Indeed, the term of order $m^{-1/2}$ shows that c is small, and the correction becomes superfluous (i.e., $c = 0$) as soon as m reaches the aforementioned value $(\gamma r \tau u_\delta)^2/\varepsilon^2$. Incidentally, note that, unlike the other quantities involved, τ is not yet known during Phase I. Hence, the bound on P_{Exc} is valid as long as the actually observed values of τ after Phase I are at most as large as the one used beforehand in obtaining m and c.

References

[1] W. H. Woodall, "The use of control charts in health-care and public-health surveillance," *Journal of Quality Technology*, vol. 38, no. 2, pp. 89–104, 2006.

[2] J. Thor, J. Lundberg, J. Ask et al., "Application of statistical process control in healthcare improvement: systematic review," *Quality and Safety in Health Care*, vol. 16, no. 5, pp. 387–399, 2007.

[3] S. H. Shaha, "Acuity systems and control charting," *Quality management in health care*, vol. 3, no. 3, pp. 22–30, 1995.

[4] C. Sonesson and D. Bock, "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society. Series A*, vol. 166, no. 1, pp. 5–21, 2003.

[5] J. Y. Liu, M. Xie, T. N. Goh, and P. Ranjan, "Time-between-events charts for on-line process monitoring," in *Proceedings of the IEEE International Engineering Management Conference (IEMC '04)*, pp. 1061–1065, October 2004.

[6] Z. Yang, M. Xie, V. Kuralmani, and K. L. Tsui, "On the performance of geometric charts with estimated control limits," *Journal of Quality Technology*, vol. 34, no. 4, pp. 448–458, 2002.

[7] M. Xie, T. N. Goh, and X. S. Lu, "A comparative study of *CCC* and *CUSUM* charts," *Quality and Reliability Engineering International*, vol. 14, no. 5, pp. 339–345, 1998.

[8] H. Ohta, E. Kusukawa, and A. Rahim, "A *CCC-r* chart for high-yield processes," *Quality and Reliability Engineering International*, vol. 17, no. 6, pp. 439–446, 2001.

[9] Z. Wu, X. Zhang, and S. H. Yeo, "Design of the sum-of-conforming-run-length control charts," *European Journal of Operational Research*, vol. 132, no. 1, pp. 187–196, 2001.

[10] W. Albers, "The optimal choice of negative binomial charts for monitoring high-quality processes," *Journal of Statistical Planning and Inference*, vol. 140, no. 1, pp. 214–225, 2010.

[11] W. Albers, "Control charts for health care monitoring under overdispersion," *Metrika*, vol. 74, no. 1, pp. 67–83, 2011.

[12] K. Poortema, "On modelling overdispersion of counts," *Statistica Neerlandica*, vol. 53, no. 1, pp. 5–20, 1999.

[13] A. Christensen, H. Melgaard, J. Iwersen, and P. Thyregod, "Environmental monitoring based on a hierarchical Poisson-gamma model," *Journal of Quality Technology*, vol. 35, no. 3, pp. 275–285, 2003.

[14] Y. Fang, "*c*-charts, *X*-charts, and the Katz family of distributions," *Journal of Quality Technology*, vol. 35, no. 1, pp. 104–114, 2003.

[15] S. H. Steiner, R. J. Cook, V. T. Farewell, and T. Treasure, "Monitoring surgical performance using riskadjusted cumulative sum charts," *Biostatistics*, vol. 1, no. 4, pp. 441–452, 2000.

[16] O. Grigg and V. Farewell, "An overview of risk-adjusted charts," *Journal of the Royal Statistical Society. Series A*, vol. 167, no. 3, pp. 523–539, 2004.

[17] O. A. Grigg and V. T. Farewell, "A risk-adjusted sets method for monitoring adverse medical outcomes," *Statistics in Medicine*, vol. 23, no. 10, pp. 1593–1602, 2004.

[18] R. Chen, "A surveillance system for congenital malformations," *Journal of the American Statistical Association*, vol. 73, pp. 323–327, 1978.

[19] R. Chen, R. R. Connelly, and N. Mantel, "The efficiency of the sets and the cuscore techniques under biased baseline rates," *Statistics in Medicine*, vol. 16, no. 12, pp. 1401–1411, 1997.

[20] W. Albers and W. C. M. Kallenberg, "Estimation in Shewhart control charts: effects and corrections," *Metrika*, vol. 59, no. 3, pp. 207–234, 2004.

[21] W. Albers and W. C. M. Kallenberg, "Are estimated control charts in control?" *Statistics*, vol. 38, no. 1, pp. 67–79, 2004.

[22] W. Albers, W. C. M. Kallenberg, and S. Nurdiati, "Parametric control charts," *Journal of Statistical Planning and Inference*, vol. 124, no. 1, pp. 159–184, 2004.

[23] W. Albers, W. C. M. Kallenberg, and S. Nurdiati, "Exceedance probabilities for parametric control charts," *Statistics*, vol. 39, no. 5, pp. 429–443, 2005.

[24] M. Riaz, "Monitoring process mean level using auxiliary information," *Statistica Neerlandica*, vol. 62, no. 4, pp. 458–481, 2008.

[25] E. L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York, NY, USA, 1959.