Theoretical and Empirical Efficiency of Sampling Strategies for Estimating Upper Arm Elevation

PER LIV1,2, SVEND ERIK MATHIASSEN1* and SUSANNE WULFF SVENDSEN3

1Department of Occupational and Public Health Sciences, Centre for Musculoskeletal Research, University of Gävle, SE-801 76 Gävle, Sweden; 2Department of Public Health and Clinical Medicine, Umeå University, SE-901 87 Umeå, Sweden; 3Department of Occupational Medicine, Danish Ramazzini Center, Herning Hospital, DK-7400 Herning, Denmark

Received 7 July 2010; in final form 19 October 2010

Objectives: To investigate the statistical efficiency of strategies for sampling upper arm elevation data, which differed with respect to sample sizes and sample allocations within and across measurement days. The study was also designed to compare standard theoretical predictions of sampling efficiency, which rely on several assumptions about the data structure, with “true” efficiency as determined by bootstrap simulations.

Methods: Sixty-five sampling strategies were investigated using a data set containing minute-by-minute values of average right upper arm elevation, percentage of time with an arm elevated <15°, and percentage of time with an arm elevated >90° in a population of 23 house painters, 23 car mechanics, and 26 machinists, all followed for four full working days. Total sample times per subject between 30 and 240 min were subdivided into continuous time blocks between 1 and 240 min long, allocated to 1 or 4 days per subject. Within day(s), blocks were distributed using either a random or a fixed-interval principle. Sampling efficiency was expressed in terms of the variance of estimated mean exposure values of 20 subjects and assessed using standard theoretical models assuming independence between variables and homoscedasticity. Theoretical performance was compared to empirical efficiencies obtained by a nonparametric bootstrapping procedure.

Results: We found the assumptions of independence and homoscedasticity in the theoretical model to be violated, most notably expressed through an autocorrelation between measurement units within working days. The empirical variance of the mean exposure estimates decreased, i.e. sampling efficiency increased, for sampling strategies where measurements were distributed widely across the day. Thus, the most efficient allocation strategy was to organize a sample into 1-min block collected at fixed time intervals across 4 days. Theoretical estimates of efficiency generally agreed with empirical variances if the sample was allocated into small blocks, while for larger block sizes, the empirical ‘true’ variance was considerably larger than predicted by theory. Theory overestimated efficiency in particular for strategies with short total sample times per subject.

Conclusions: This study demonstrates that when exposure data are autocorrelated within days—which we argue is the major reason why theory overestimates sampling performance—sampling efficiency can be improved by distributing the sample widely across the day or across days, preferably using a fixed-interval strategy. While this guidance is particularly valid when small proportions of working days are assessed, we generally recommend collecting more data than suggested by theory if a certain precision of the resulting exposure estimate is needed. More data per se give a better precision and sampling larger proportion(s) of the working day(s) also alleviate the negative effects of possible autocorrelation in data.

Keywords: arm elevation; exposure assessment; precision; sample allocation; sampling strategy; statistical efficiency

*Author to whom correspondence should be addressed. Tel: +46-26-648206; fax: +46-26-648686; e-mail: smnc@hig.se
INTRODUCTION

Valid and reliable exposure data are essential in both epidemiological and intervention studies, as well as in surveillance of the work environment for the purpose of research and ergonomics practice. Insufficient exposure data have been pointed out by several authors as one reason why many studies have failed to find clear relations between exposure and disorders (Armstrong et al., 1993; Winkel and Mathiassen, 1994; van der Beek and Frings-Dresen, 1998; Burdorf and van der Beek, 1999; Punnett and Wegman, 2004). Much effort has been put into developing instruments, either technical measurement devices, observation-based methods, or self-reports, which yield unbiased exposure estimates; that is, results that will be valid in the long run (Wiktorin et al., 1994; van der Beek and Frings-Dresen, 1998; Burdorf and van der Beek, 1999; Punnett and Wegman, 2004). However, exposure varies between individuals, between working days within individuals, and within working days. Due to this variability, even exposure estimates obtained by unbiased instruments are inherently uncertain because they have to be based on samples of subjects and time periods. This presents the researcher or practitioner with the question of how to design an appropriate measurement strategy suit the specific purpose of the investigation. Even in the simple case of documenting the mean exposure in a group, decisions are needed on an appropriate number of subjects and a sufficient number of measurements from each subject. While studies based on whole-day measurements of biomechanical exposures can be found in the literature (Svendsen et al., 2004a; Leijon et al., 2005; Hansson et al., 2009, 2010), the majority of studies have measured only parts of working day(s). In these cases, decisions also have to be made about a proper allocation in time of measurements within day(s). A general aim which should guide these decisions is to secure a high efficiency, that is, to obtain as much information as possible at a given investment of resources, or equivalently, to arrive at a desired level of performance with as little measurement effort as possible.

Statistical models are available that predict the performance—with respect to the precision of the resulting exposure estimate—of a particular measurement strategy on the basis of the size of the exposure sample and the structure of exposure variability, in terms of variance components (Samuels et al., 1985; Searle et al., 2006). While statistical guidelines for efficient design of measurement strategies have been discussed extensively within the field of chemical exposure assessment (Samuels et al., 1985; Rappaport et al., 1995; Tielemans et al., 2002; Loomis and Kromhout, 2004; Lampa et al., 2006; Chen et al., 2009), the literature addressing biomechanical exposures is limited. A few studies have determined variance components in exposure between and within individuals and demonstrated how to use them for designing measurement strategies (Mathiassen et al., 2002, 2003a,b; Nordander et al., 2004); methodological sources of variance have been included in rare cases (Jackson et al., 2009). However, the statistical model used in these applications is based on a number of assumptions about the data. For example, subjects, days within subjects, and measurements within days are assumed to be statistically independent, and variances between measurement units are assumed to be equal at each of these levels. These assumptions are probably rarely met for biomechanical exposure data, as suggested by a few reports (Mathiassen et al., 2003a). Partly in response to this fallacy, distribution-free bootstrapping techniques (Efron and Tibshirani, 1993) have been developed that determine performance on the basis of the empirical characteristics of the data. Some studies of biomechanical exposure have used these techniques for exploring measurement strategies (Burdorf and van Riel, 1996; Hoozemans et al., 2001; van Dieen et al., 2002; Paquet et al., 2005; Fethke et al., 2007; Mathiassen and Paquet, 2010). However, little research has examined deviations of biomechanical exposure data from the assumptions associated with analytical assessments of measurement strategies, and no studies have, to our knowledge, been devoted to evaluating the effects that such deviations might have on sampling performance.

Thus, the primary purpose of this study was to examine the effects of total sample time per subject and allocation of samples in time on the performance of sampling strategies for assessing three upper arm elevation variables in groups. A second purpose was to compare ‘true’ empirical sampling performance with the performance predicted by analytical theory, which rely on a number of assumptions about the structure of data.

MATERIALS AND METHODS

Study population and exposure measurements

Data were obtained from a previously published epidemiological study of the relationship between upper arm elevation and shoulder disorders (Svendsen et al., 2004a,b, 2005). In that study, upper arm elevation was recorded for entire working days among machinists, car mechanists, and house painters. Within a defined geographical area, all
relevant companies with five or more journeymen were identified in the Danish Central Business Register. This resulted in a list of 29 machine shops, 110 garages for domestic cars, and 119 painter’s workshops, employing a total of 942 machinists, 692 car mechanics, and 1579 house painters. From the companies, 13 pairs of colleagues were sampled randomly for each occupational group, i.e. 26 subjects per group. Selected subjects were included in the study if they had at least 1 year of employment, were males 30–65 years of age, and had at least four scheduled working days in the planned measurement week. Subjects were excluded if they had shoulder complaints that interfered with their performance at work. If an initially selected subject was excluded or unwilling to participate, he was replaced by another randomly sampled subject, preferably from the same company.

For each subject, right and left upper arm elevation was recorded for all days in a week using the Abduflex system (Mortimer et al., 1999). The Abduflex registered upper arm angle with respect to the line of gravity in six 15° intervals from 0 to 90° and a seventh interval for all angles >90°. Data were collected to a portable logger at a frequency of 1 Hz and downloaded at the end of the day to a laptop.

Data processing

The present study only used data from the right arm. All working days comprising <240 min of data were excluded from the data set. Thereafter, subjects with less than four accepted working days were excluded from further analysis. At this point, the study population contained 23 housepainters, 23 car mechanics, and 26 machinists. For each of these subjects, four working days were selected at random. The duration of the recordings ranged between 240 and 721 min, while most of them were close to 480 min. Sixty-two of all 288 days were <420 min, and only 6 days were >480 min.

For all accepted days, minute-by-minute values were obtained for three variables characterizing arm elevation during work: ‘average elevation’, as a general descriptive variable, the percentage of time spent with an arm elevation >90° (‘percentage time >90°’), which was shown to be a risk factor for developing shoulder disorders in the original epidemiological study (Svendsen et al., 2004a) and the percentage of time with an arm elevation <15° (‘percentage time <15°’), to indicate the occurrence of neutral postures.

While the two latter variables could be directly assessed from the categorical data recorded by the Abduflex, average elevation was estimated by first replacing data in each of the Abduflex intervals by the central angle value for that interval, except for the seventh category, i.e. angles >90°, which was assigned the value 105° and then using this transformed data as the basis for estimating the mean. Occasional short periods of missing data (in all 299 min of a total of 126 824 min of recording) were linearly interpolated.

In order to achieve a balanced data set, all working days were processed to contain exactly 480 min of data. The six working days exceeding 480 min were truncated at 480 min, and for all working days <480 min, new data were added. For a specific working day, data to be added were obtained using a moving block bootstrap procedure (Efron and Tibshirani, 1993), identifying new data by selecting 30-min data blocks with replacement from the available data of that day. Of a total of 138 240 min of data after this procedure, 11 416 min were simulated. The new, partly simulated, ‘parent’ data set was shown to have similar variance components and autocorrelation (see below) as the original unbalanced data set. In order to further check the validity of the simulation, all original working days >440 min were selected. These ‘long’ working days were truncated at 240 min and padded using the procedure described above. These partly simulated working days, 480 min long, were then compared to the original working days. Only marginal differences in exposure levels and autocorrelation (see below) were found, indicating that the procedure of adding data by resampling preserved the structure and characteristics of the original data set. As an illustration, Fig. 1 shows minute-by-minute values for the three posture variables from a working day for a car mechanic.

Sampling strategies

Using the described parent data set, 65 different sampling strategies were investigated. Strategies were designed to explore how efficiency depends on sample size, allocation of samples between days, and allocation of samples within days (Table 1). All strategies involved sampling of data from 20 subjects, representing a realistic size of a field study assessing the exposure of a population using technical measurements or observations. A specific sampling strategy was constructed by first selecting the total sample time per subject. Four sizes of this sample time were investigated: 30, 60, 120, and 240 min. For each of these four values of sample time, effects of splitting the sample into continuous periods (blocks) of $t_b = 1, 5, 15, 30, 60, 120,$ or 240 min were examined. For example, if a sample time of 240 min was to be allocated to blocks of $t_b = 30$ min, the sample
was split into 240/30 = 8 different blocks of 30 min each.

Next, the sampling blocks were allocated to either 1 or 4 days, evenly if possible. Thus, allocating eight blocks across 4 days would lead to two blocks being sampled per day. Excess blocks that could not be evenly distributed were randomly allocated to days. For example, when 30 min were sampled in 5-min blocks from 4 days, two random days received two blocks, while only one block was allocated to each of the remaining 2 days.

The resulting blocks were then distributed within day(s) at fixed intervals or randomly. In fixed-interval sampling, the blocks were distributed evenly across the day, i.e. the time interval between each block had a fixed value. In random sampling, the time intervals between blocks were randomly selected, constrained on their sum being equal to the total non-sampled time in the investigated strategy. While Table 1 suggests that 112 strategies (i.e. 4 × 2 × 7 × 2) would be investigated, several of these combinations are nonexistent. For instance, the total sample time per subject cannot be smaller than the block size. Furthermore, in the case of sampling only one block per day, fixed-interval sampling is equivalent to random sampling and therefore a redundant strategy. Leaving out impossible and redundant combinations led to 65 investigated sampling strategies.

### Theoretical sampling efficiency

The necessary information on exposure variability for estimating the theoretical efficiency of each strategy was obtained by analyzing data using a standard two-way random effects model (Loomis and Kromhout, 2004; Burdorf, 2005):

\[
y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{k(ij)},
\]

where \(i = 1, 2, \ldots, a; j = 1, 2, 3, 4; k = 1, 2, \ldots, n\). \(y_{ijk}\) is the \(k\)th observation of exposure in subject \(i\) on working day \(j\), \(\mu\) is the grand mean of the exposure, \(\alpha_i\) is the random effect on exposure of the \(i\)th subject, \(\beta_{j(i)}\) is the random effect of day \(j\) within subject \(i\), and \(\epsilon_{k(ij)}\) is the random error for observation \(k\) within day \(j\) for subject \(i\). The model assumes that the variables \(\alpha_i, \beta_{j(i)},\) and \(\epsilon_{k(ij)}\) are mutually independent and identically distributed with zero mean and equal variances between subjects, between days, and within days—\(\sigma_{BS}^2, \sigma_{BD}^2,\) and \(\sigma_{WD}^2\), respectively—for all \(i, j,\) and \(k\) (i.e. homoscedasticity). In the present study, the variance components \(\sigma_{BS}^2, \sigma_{BD}^2,\) and \(\sigma_{WD}^2\) were estimated in the parent data set for each of the three occupational groups and for each of the three posture variables using restricted maximum likelihood (REML) (Searle et al., 2006). Under this model, the variance of an estimated arithmetic mean, \(\sigma_{\text{theory}}^2\), based on a sample of \(n_\text{q}\) 1-min observation from each of \(n_\text{d}\) 480 min long working days in

---

### Table 1. Levels of the parameters determining the investigated sampling strategies

<table>
<thead>
<tr>
<th>Sample time per subject (min)</th>
<th>30, 60, 120, 240</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of days, (n_\text{d})</td>
<td>1, 4</td>
</tr>
<tr>
<td>Block size (t_b) (min)</td>
<td>1, 5, 15, 30, 60, 120, 240</td>
</tr>
<tr>
<td>Regularity within days</td>
<td>Random or fixed interval</td>
</tr>
</tbody>
</table>
predicted value of \( \beta \) for all three predicted values were calculated. Furthermore, the predicted values of \( \beta_{ij} \) were checked by Levene’s test of equal variance at a 0.05 significance level. The autocorrelation function for measurement units within days was estimated for lags 1, 2, 3, 4, 5, and 10 min (Francis et al., 1989).

Empirical sampling efficiency

A nonparametric bootstrapping technique (Efron and Tibshirani, 1993) was used to estimate the empirical efficiency of each sampling strategy. The bootstrap procedure used here was similar to the cluster bootstrap suggested by Field and Welsh (2007), although modified for a hierarchical data set with three levels: i.e. subjects, days, and measurement units within days. It was practiced as follows for a particular sampling strategy: first, 20 subjects were selected with replacement from the parent data set. Second, from each selected subject, \( n_d \) days were selected without replacement. Third, from the recordings of the selected days, the number and temporal location of sampling blocks were identified using the procedures described above. Sampling within days was performed without replacement to ensure that the temporal structure of the data within days was kept intact in the simulation. This whole procedure created an offspring bootstrapped data set of the size and structure specified by the sampling strategy. The group means of the three posture variables were then calculated for the offspring dataset. For each sampling strategy, this bootstrap procedure was repeated 10,000 times and the sample variance, \( s^2_{\text{bootstrap}} \), of the 10,000 group mean values of the three posture variables was used as the empirical bootstrap estimate of the efficiency of that particular strategy. Ten thousand bootstrap replicates were considered sufficient to give stable estimations of empirical variance (Efron and Tibshirani, 1993). Programming was done in MATLAB 7.0.4 (The MathWorks, Inc.) and in R version 2.9.2 (R Development Core Team, 2009).

Design effect

Empirical and theoretical efficiencies of a particular sampling strategy were compared using the design effect factor (deff) (Mathiassen et al., 2003a):

\[
deLeff = \frac{s^2_{\text{bootstrap}}}{s^2_{\text{theory}}}
\]

The deff of a sampling strategy measures how correctly theory predicts its ‘true’ empirical efficiency. A deff equal to 1 implies that the theoretical model results in a correct estimate, while deff values larger or smaller than 1 imply that the actual precision of a mean exposure estimate is worse or better, respectively, than predicted by theory. By expressing efficiency in terms of (relative) deff values, sampling performance can be compared across different variables and settings. This is not readily possible on the basis of absolute values of variance.

RESULTS

Exposure distributions

Table 2 gives descriptive statistics of the data sets, including REML estimates of variance components. Table 2 also reports estimates of the autocorrelation function of the data for lags 1, 2, 3, 4, 5, and 10 min. The autocorrelation was estimated to be well over zero in all cases, confirming the impression from Fig. 1; i.e. values close in time are similar. This implies a violation against the assumption of independence between random errors in the theoretical models. Furthermore, Table 2 presents skewness and kurtosis of the predictors \( z_i, \beta_{ij}, \) and \( \varepsilon_{ij} \) for each occupational group and posture variable. A normal distribution has a skewness of 0 and a kurtosis of 3. Homoscedasticity for between-days variance components was rejected by Levene’s test in all nine cases, except for average elevation among machinists. Levene’s tests were not performed on the predictors of \( \varepsilon_{ij} \) because they were too many to allow for formal hypothesis testing. However, graphs of the predictors of \( \varepsilon_{ij} \) revealed that heteroscedasticity was clearly present even at that level.

Empirical versus theoretical sampling performance

Deff values from all strategies, occupational groups, and posture variables are presented in Table 3. In most
cases, deff was larger than 1, showing that the ‘true’
empirical variance was larger than that predicted by
theory. For strategies using a block size of 1, as well
as for a large sample time per subject in general, the
deffs were, however, close to 1, indicating that theory
predicted efficiency well.

While the deff measures the relationship between
empirical and theoretical variance, absolute values
of performance can be obtained by simple arithmetic.
For instance, the deff corresponding to 60 min
of random sampling from one working day using
a block size of 5 was 1.44 for ‘percentage time
%90; among car mechanics. According to Table 2,
the variances between subjects, between days, and
within days for this variable were 3.0, 4.0, and
164.7, respectively. Using equation (2), the
theoretical variance of the mean value of 20 car
mechanics, each approached for 60 min during
1 day is

\[
\sigma^2_{\text{theory}} = \frac{3.0}{20} + \frac{4.0}{20 \times 1} + \left(1 - \frac{60}{480}\right)\frac{164.7}{20 \times 1 \times 60} = 0.47.
\]

Given a deff of 1.44, the empirical variance,
\(\sigma^2_{\text{bootstrap}}\), was 0.47 \times 1.44 = 0.68.

Effects of the sample size

In theory, increasing the number of sampled
minutes within each measurement day will lead to
less variance of the estimated mean [equation (2)].
This effect was confirmed by the empirical results
(Table 3), and for block sizes larger than one, empirical
variance was even reduced at a faster rate than
predicted by theory, approaching the theoretical
variance at large sample sizes. An example of this
effect is illustrated in Table 4.

Allocation between versus within days

In theory, the precision of an estimated exposure
mean will increase if a certain total number of measure-
ment units from a certain number of subjects
are distributed across more days [equation (2)].
This effect was confirmed by the bootstrap variances,
but the gain in precision when distributing a certain sam-
ples across 4 days rather than 1 was less than predicted
by theory, as shown by the deffs for the 4 days alloca-
tions being generally larger than those for the corre-
sponding 1-day allocation. As an example (Table 4),
random sampling from 20 car mechanics of 60 min
of data on percentage time \(\%90;\), allocated in 5-min
block within 1 day, led to a theoretical variance of
0.47 and an empirical variance of 0.68. Distributing
Table 3. Empirical versus theoretical variance of all investigated strategies, as expressed by the deff ratio [equation (3)].  

<table>
<thead>
<tr>
<th>Machinists, average elevation, 1 day</th>
<th>Car mechanics, average elevation, 1 day</th>
<th>House painters, average elevation, 1 day</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1/n_0 = 30$</td>
<td>$n_1/n_0 = 60$</td>
<td>$n_1/n_0 = 120$</td>
</tr>
<tr>
<td>$t_0$</td>
<td>$t_0$</td>
<td>$t_0$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>240</td>
<td>240</td>
<td>240</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Machinists, %&lt;15°, 1 day</th>
<th>Car mechanics, %&lt;15°, 1 day</th>
<th>House painters, %&lt;15°, 1 day</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1/n_0 = 30$</td>
<td>$n_1/n_0 = 60$</td>
<td>$n_1/n_0 = 120$</td>
</tr>
<tr>
<td>$t_0$</td>
<td>$t_0$</td>
<td>$t_0$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>240</td>
<td>240</td>
<td>240</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Machinists, %&gt;90°, 1 day</th>
<th>Car mechanics, %&gt;90°, 1 day</th>
<th>House painters, %&gt;90°, 1 day</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1/n_0 = 30$</td>
<td>$n_1/n_0 = 60$</td>
<td>$n_1/n_0 = 120$</td>
</tr>
<tr>
<td>$t_0$</td>
<td>$t_0$</td>
<td>$t_0$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>240</td>
<td>240</td>
<td>240</td>
</tr>
</tbody>
</table>
Table 3. Efficiency of sampling strategies for estimating upper arm elevation

<table>
<thead>
<tr>
<th>Machinists, average elevation, 4 days</th>
<th>Car mechanics, average elevation, 4 days</th>
<th>House painters, average elevation, 4 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_s n_i = 30$</td>
<td>$n_s n_i = 30$</td>
<td>$n_s n_i = 30$</td>
</tr>
<tr>
<td>$n_s n_i = 60$</td>
<td>$n_s n_i = 60$</td>
<td>$n_s n_i = 60$</td>
</tr>
<tr>
<td>$n_s n_i = 120$</td>
<td>$n_s n_i = 120$</td>
<td>$n_s n_i = 120$</td>
</tr>
<tr>
<td>$n_s n_i = 240$</td>
<td>$n_s n_i = 240$</td>
<td>$n_s n_i = 240$</td>
</tr>
<tr>
<td>$t_s$</td>
<td>$r$</td>
<td>$f$</td>
</tr>
<tr>
<td>1</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>5</td>
<td>1.28</td>
<td>1.31</td>
</tr>
<tr>
<td>15</td>
<td>1.30</td>
<td>1.12</td>
</tr>
<tr>
<td>30</td>
<td>1.21</td>
<td>1.02</td>
</tr>
<tr>
<td>60</td>
<td>1.06</td>
<td>1.05</td>
</tr>
<tr>
<td>120</td>
<td>1.07</td>
<td>1.06</td>
</tr>
<tr>
<td>240</td>
<td> </td>
<td> </td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Machinists, %&lt;15°, 4 days</th>
<th>Car mechanics, %&lt;15°, 4 days</th>
<th>House painters, %&lt;15°, 4 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_s n_i = 30$</td>
<td>$n_s n_i = 30$</td>
<td>$n_s n_i = 30$</td>
</tr>
<tr>
<td>$n_s n_i = 60$</td>
<td>$n_s n_i = 60$</td>
<td>$n_s n_i = 60$</td>
</tr>
<tr>
<td>$n_s n_i = 120$</td>
<td>$n_s n_i = 120$</td>
<td>$n_s n_i = 120$</td>
</tr>
<tr>
<td>$n_s n_i = 240$</td>
<td>$n_s n_i = 240$</td>
<td>$n_s n_i = 240$</td>
</tr>
<tr>
<td>$t_s$</td>
<td>$r$</td>
<td>$f$</td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>1.16</td>
<td>1.19</td>
</tr>
<tr>
<td>15</td>
<td>1.17</td>
<td>1.06</td>
</tr>
<tr>
<td>30</td>
<td>1.12</td>
<td>1.05</td>
</tr>
<tr>
<td>60</td>
<td>1.00</td>
<td>1.01</td>
</tr>
<tr>
<td>120</td>
<td> </td>
<td> </td>
</tr>
<tr>
<td>240</td>
<td> </td>
<td> </td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Machinists, %&gt;90°, 4 days</th>
<th>Car mechanics, %&gt;90°, 4 days</th>
<th>House painters, %&gt;90°, 4 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_s n_i = 30$</td>
<td>$n_s n_i = 30$</td>
<td>$n_s n_i = 30$</td>
</tr>
<tr>
<td>$n_s n_i = 60$</td>
<td>$n_s n_i = 60$</td>
<td>$n_s n_i = 60$</td>
</tr>
<tr>
<td>$n_s n_i = 120$</td>
<td>$n_s n_i = 120$</td>
<td>$n_s n_i = 120$</td>
</tr>
<tr>
<td>$n_s n_i = 240$</td>
<td>$n_s n_i = 240$</td>
<td>$n_s n_i = 240$</td>
</tr>
<tr>
<td>$t_s$</td>
<td>$r$</td>
<td>$f$</td>
</tr>
<tr>
<td>1</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>1.78</td>
<td>1.30</td>
</tr>
<tr>
<td>15</td>
<td>1.37</td>
<td>1.41</td>
</tr>
<tr>
<td>30</td>
<td>1.41</td>
<td>1.22</td>
</tr>
<tr>
<td>60</td>
<td>1.22</td>
<td>1.23</td>
</tr>
<tr>
<td>120</td>
<td>1.29</td>
<td>1.28</td>
</tr>
<tr>
<td>240</td>
<td>1.31</td>
<td>1.31</td>
</tr>
</tbody>
</table>
the same total sample time across 4 days, showing a deff of 1.67 (Table 3), gave theoretical and empirical variances of 0.33 and 0.55, respectively.

Effects of the sample block size
For a certain sample time per subject, the deff increased with block size (Table 3). Since the theoretical variance does not change with block size, this result shows that ‘true’ empirical precision decreases with larger blocks. This effect was particularly prominent for small sample sizes per subject. Figure 2 illustrates this interaction between block size and total sample size per subject for car mechanics, ‘percentage time >90°’, random sampling during 1 day. ‘Percentage time >90°’ showed the most pronounced effect of block size, while variance increased only moderately with block size for the variable ‘percentage time <15°’ (Table 3).

Random versus fixed-interval sampling
In general, fixed-interval sampling performed better than random sampling, yet the difference was small (Table 3). In a few cases, fixed-interval sampling even led to a slightly larger variance than random sampling.

DISCUSSION
Our results suggested that estimates of a group mean value based on samples from parts of working days are, in general, considerably less precise than predicted by sampling theory. This is particularly evident when smaller proportions of working days are sampled and when subsamples within days are collected close to each other in time. Thus, if a certain precision of the estimate is required, for instance, to ensure sufficient power in an ergonomics intervention study, more data need to be collected than what appears from standard analytical guidelines. Also, the size of standard confidence intervals for mean exposure estimates may be underestimated, in particular for single individuals assessed for short periods of time. Standard statistical guidelines for sampling are based on a number of assumptions regarding exposure distributions, which were shown to be violated in the present material; autocorrelation and heteroscedasticity occurred in several cases.

While our data were collected in only three occupational groups, the diversity in exposure level and variability among those groups suggest that our general results apply to a variety of trades, including routinized (exemplified by the machinists) as well as less routinized jobs (exemplified by the house painters). This includes the positive effect of distributing samples across 4 days instead of 1, which was, as expected, more prominent among the
house painters, who showed comparatively large between-days exposure variability (cf. Table 2).
Also, the finding that precision of mean exposure according to theory was better than empirical precision came out in all three groups, in spite of their profound dissimilarities in exposure characteristics. This effect, particularly prominent for small samples close in time, was not clearly related to exposure level or size of variance components.

**Bootstrap estimates and the resulting design effects**

In the present study, we considered the variances estimated by bootstrapping to be the ‘truth’ against which theoretical variances should be compared.

We have good reasons to believe that bootstrap variance estimates are much closer to the ‘truth’ than theoretical estimates, which rely on assumptions that were obviously invalid in our material. However, the validity of the bootstrap is also uncertain to some extent. We performed bootstrapping on a hierarchical data set, consisting of subject, days, and measurement units within days. Various methods for bootstrapping hierarchical data have been suggested (McCullagh, 2000; Carpenter et al., 2003; Field and Welsh, 2007). However, all hierarchical bootstrap methods in the literature deal with two-level data sets, while our data have a three-level structure. To our knowledge, no theoretical discussion has been published of using the bootstrap on hierarchical data sets with more than two levels. However, the results from random sampling with block size one in the present study can be used to examine the validity of the used three-level bootstrap. Sampling at random using a block size of 1 implies that the assumption of independent errors in the theoretical model [equation (1)] cannot be violated. Thus, the deff should be very close to 1 for these strategies provided that autocorrelation in data is the major explanation of empirical variances deviating from theory (see below). However, we found deffs in most of these cases to be slightly < 1 (Table 3). This finding might be explained by two effects: (i) the variance components used in the theoretical calculations were biased in being estimated from autocorrelated data, which leads to a slightly upward biased theoretical estimation and/or (ii) the bootstrap estimate of the variance was slightly downward biased. However, since deviations from a deff of 1 in random samplings using 1-min block were small and consistent, we believe that the three-level bootstrap generated estimations that were sufficiently correct for valid conclusions to be drawn from our study.

Two of the investigated 4-day sampling strategies led to an unbalanced data set, in the sense that samples could not be evenly distributed across days: a 30-min sample allocated in 1-min blocks and 5-min blocks. In these two cases, the theoretical variance [equation (2)] was calculated assuming a balanced allocation, i.e. 7.5 min of data collected from each day. This alleged balancing of the theoretical data will lead to an additional small downward bias of their estimated variance compared to an unbalanced allocation (Samuels et al., 1985), but the effect was too small to be detectable in the deff values (Table 3).

**Data processing**

Data were delivered from the Abduflex device at 1 Hz and processed by us into minute-by-minute values of upper arm posture. This resolution was selected in order to generate a manageable data set, while still capturing exposure variability within days. However, since autocorrelation was present in the data, averaging affects exposure variability differently from what is expected by theory. Thus, the relationships between empirical and theoretical efficiency obtained in the present study are only strictly valid for data sets using 1 min as the basic measurement unit.

When collecting data in field studies, it is unlikely to achieve a balanced data set where all measured working days for all workers contain exactly the same quantity of data. Unbalancedness could be caused by, for instance, loss of data due to technical problems or monitored work shifts differing in duration. The precision of a mean exposure estimate based on a certain sample size within a day will depend on the total amount of data available for that day and so is sensitive to the total measurement duration of the day. Some dispersion in the amount of data collected per day was, indeed, present in our material. We aimed at obtaining sampling performance results that could be generalized to other occupational groups with a similar exposure structure and therefore decided to process the original data into a balanced set, where all working days comprised exactly 480 min of data. Since we found the partly simulated days to differ only marginally in exposure characteristics from the original, true days, we believe that the data manipulation did not influence the final results to any noticeable extent.

**Group size**

We examined the precision of each sampling strategy in terms of the variance of a mean exposure of 20 subjects. According to equation (2), changing the number of subjects in a sample will scale the variances of all allocation strategies within subjects to the same extent. The theoretical estimation and the bootstrap procedure assume that subjects are sampled at random. Therefore, it is reasonable to believe that
using other numbers than 20 would have resulted in similar effects on both theoretical and bootstrap variances and so that the deffs (Table 3) would apply irrespective of group size.

**Exposure levels and variability in the three groups**

As expected and intended in the design of the original study (Svendsen et al., 2004a,b, 2005), the three occupational groups showed markedly different exposure characteristics. The house painters exhibited a considerable occurrence of work with elevated arms, the car mechanics somewhat less, whereas elevated arms rarely occurred among the machinists. Due to the selection of different posture variables, our results cannot be directly compared with the majority of the literature of upper arm postures in occupational groups. However, in a study of hairdressers, Veiersted et al. (2008) found the percentage of time with the arms elevated >90° to be 3.0% for the right arm, as compared to 8.7% for the house painters in the present study, 4.7% for the car mechanics, and 1.6% for the machinists. Other studies have shown proportions of upper arm elevation >90° to be ~12% among construction workers (Paquet et al., 2006) and <1% among office workers (Fernström and Ericson, 1996). While variance components for upper arm elevation variables have been reported in some studies (e.g. Mathiassen et al., 2003b; Hansson et al., 2006; Wahlström et al., 2010), we did not find any reference data on the three present variables.

**Measurement strategies in the literature**

The 65 investigated sampling strategies in this study were intended to cover a wide range of allocation scenarios, in particular within working days. Previous studies quantifying exposure by technical measurements have typically sampled data in one continuous block, mainly for practical reasons (e.g. Sporrong et al., 1999; Veiersted et al., 2008), while splitting a measurement into several smaller periods within the same working day(s) has occurred mostly in studies based on observations and often in combination with fixed-interval sampling (e.g. Karhu et al., 1981; Buchholz et al., 1996). Studies discussing the statistical efficiency of different measurement strategies have mainly addressed the influence of the sample size (Burdorf and van Riel, 1996; Hoozemans et al., 2001; Mathiassen et al., 2002, 2003b; Paquet et al., 2005; Fethke et al., 2007; Trask et al., 2008; Mathiassen and Paquet, 2010). Some studies have investigated the tradeoff between selecting more subjects and more repeated measurements from each subject (Hoozemans et al., 2001). Very few studies have addressed the allocation of measurements within working days. Mathiassen et al. (2003a) investigated the performance of some within-day sampling schemes in a study of electromyography recordings from the upper trapezius muscle in cleaners and office workers. In all investigated sampling strategies, one fifth of the working day was sampled, and continuous sampling in one block was compared to sampling using a block size of 1 minute, either in a fixed-interval scheme or distributed at random. The study showed similar results to the present study; fixed-interval sampling was more efficient than random sampling and consecutive sampling led to a dramatic loss of precision of the mean exposure estimate.

**Violated assumptions and their effects on sampling performance**

The deff value was clearly >1 for most of the investigated strategies in the present study (Table 3). This suggests that data violated the assumptions underlying the theoretical model [equation (1)] and that this departure resulted in loss of precision. For hypothesis testing using analysis of variance (ANOVA), statisticians regard the assumptions of (i) normally distributed data, (ii) homoscedasticity, and (iii) independent data to be increasingly important in the stated order (Cochran, 1947; Box, 1954a,b; Gastwirth and Rubin, 1971; Glass et al., 1972; Lissitz and Chardos, 1975; Pettitt and Siskind, 1981). The same order probably also applies to the theoretical model in this study, even though we used REML algorithms.

**Normality**

The theoretical model in equation (1) does not per se assume any particular distributional form for $x_i$, $\beta_{j|i}$, and $\nu_{k|ij}$. While solving the model using ANOVA estimations does not incur any distributional assumptions, REML procedures as used by us require normality. However, for balanced data, ANOVA estimations are equal to those obtained by REML (Swallow and Monahan, 1984). The theoretically predicted precision of a mean exposure [equation (2)] is also valid for any distribution of the random effects in the statistical model. Thus, the assumption of normally distributed data is a standard, formal assumption in a thorough estimation and hypothesis testing procedure, which could have been ignored in this study since no testing was performed. Normality was not met in most cases for $x_i$ and $\beta_{j|i}$, but the departures were, in general, moderate. For “percentage time >90°”, however, non-normality of $x_i$ and $\beta_{j|i}$ was obvious, except for $x_i$ among car
mechanics. The \( \varepsilon_{k(ij)} \) was clearly non-normally distributed, except for ‘percentage time <15°’, showing a moderate departure from normality.

**Homoscedasticity**

Our data showed obvious cases of heteroscedasticity (Table 2). For hypothesis testing using ANOVA, inequality of variances has been shown to have little effect if the data set is balanced but a larger effect with unbalanced data sets (Scheffé, 1959). Since the present simulations were based on a balanced data set, it is likely that heteroscedasticity only contributed little to the incorrectness of the theoretically predicted sampling performance.

**Independence**

In the present material, postures close in time correlated considerably, leading to autocorrelated residuals in the theoretical model. Positively autocorrelated data will lead to empirical variances being larger than those predicted by theory (Rappaport, 1991), consistent with the results in the present study. The considerable influence of block size is a clear indication of the effect of autocorrelated data within working days. Sampling using large block sizes will lead to a more autocorrelated sample than sampling in small block sizes since data in large blocks lie closer in time. Another consequence of autocorrelation is that fixed-interval will, in general, perform better than random sampling because fixed-interval sampling assures that samples are distributed widely across time. While we suggest that temporal dependence between the residual errors in the theoretical model, i.e. between individual sampling units, was the primary explanation that variance was underestimated by theory, we did not find a direct association between the size of the sample autocorrelation function and the \( \text{deff} \) size across variables and occupational groups.

Thus, sampling performance is probably influenced by additional patterns of temporal dependence not captured by the autocorrelation function. If data are collected for entire days, temporal dependence of minute-by-minute exposure within days is not an issue. However, temporal dependence may also occur between working days close in time, as demonstrated in several studies on chemical exposure (Spear et al., 1986; Francis et al., 1989; Rappaport, 1991). The statistical model [equation (1)] assumed independence between working days, but this property could not be examined in the present data material due to the limited number of available working days for each worker. Autocorrelation between days would lead to an additional downward bias of theoretically estimated sampling efficiency.

**CONCLUSIONS AND GUIDANCE**

This study showed that the theoretically predicted precision of mean estimates of variables describing upper arm elevation may, in general, be better than the ‘true’ precision, as determined by empirical methods, in particular if small proportions of working days are assessed and samples are collected close in time. This deviance is caused by violated assumptions in the statistical model underlying the theoretical estimation, mainly caused by autocorrelation between exposure measurement units within working days.

The dependency between exposure units directly reflects the structure across time of upper arm postures in individual workers. Therefore, the efficiency effects of allocating samples in different ways should be valid, regardless of whether postures were recorded by technical instruments or observations. Since we found similar results for three different posture variables in three occupational groups with very different exposure characteristics and since some of the findings are supported by previous studies of other exposure variables in other occupations (Mathiassen et al., 2003a), we believe that our results have a wide applicability.

On the basis of our study, we therefore propose the following statements for practitioners and researchers sampling biomechanical exposure data from parts of working days. This guidance adds to previous recommendations, for instance, concerning the need to decide in advance for the necessary efficiency of the data collection strategy and the need to base the planned design on pilot or literature data on exposure variability (Burdorf and van Riel, 1996; Burdorf and van der Beek, 1999; Mathiassen et al., 2002):

- We recommend collecting more data than suggested by theory. More data \textit{per se} gives a better precision of a mean exposure estimate, and sampling larger proportion(s) of the working day(s) also alleviate the negative effects of possible autocorrelation in data.
- We also recommend distributing measurements widely across time; if possible by using a fixed-interval approach during >1 day.

**FUNDING**


**REFERENCES**

Mathiassen SE, Burdorf A, van der Beek AJ et al. (2003a) Efficient one-day sampling of mechanical job exposure data—a study based on upper trapezius activity in cleaners and office workers. AIHA J (Fairfax, Va); 64: 196–211.