Assessment of Representativity of a Study Population – Experience of the Kiel Obesity Prevention Study (KOPS)

Sandra Plachta-Danielzika Carmen Bartelb Heiner Raspeb Ute Thynec Beate Landsberga

Manfred James Müller*a

aInstitut für Humanernährung und Lebensmittelkunde, Christian-Albrechts-Universität, Kiel,
bInstitut für Sozialmedizin,
cInstitut für Kinder- und Jugendheilkunde, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

Key Words
Representativity · Non-response · Missing values · Weight status

Summary
Objective: Exemplified by data of the Kiel Obesity Prevention Study (KOPS), different methods to control for response bias and to assess representativity were compared. Methods: 4,997 cross-sectional data of 5- to 7-year-old German children (main cohort) were investigated between 1996 and 2001 within school entry examination. A subgroup responded to a questionnaire to socio-demographic and lifestyle factors (responders, n = 2,631). Representativity of the main cohort was tested in comparison to the total population. To control for response bias within the responders a non-response analysis as well as an analysis of missing values were performed. Results: The comparison with the total population showed a higher prevalence of obese boys and girls from families of low socio-economic status (SES) within the main cohort. The responders were less frequently obese and overweight and more rarely belonged to low SES families when compared with non-responders. Analysis of missing values did not detect any further biases. According to an epidemiological assessment of differences the main cohort of KOPS is suggested to be representative for all 5- to 7-year-old children in Kiel, whereas the responders can be at best called ‘relatively’ representative. Conclusion: The analysis of non-response is the most sensitive method to detect group differences, but a comparison with the total population can also be used to control for biases. In addition representativity has to be proven not only for the main cohort but also for the subgroup of responders with which data analysis will be done.

Introduction
Epidemiological studies have to take into account representativity of the study population. This is a precondition to generalize the results and to compare them to other studies. Recently published reviews on prevention programs for reducing childhood overweight assessed the methodology of the studies to characterize their evidence quality [1–3]. In all methodology checklists used, external validity or selection bias was considered. In practice, most authors confirm the representativity of their study populations in one sentence without providing additional information. Only few studies have fully described procedures used to test for representativity [e.g. 4–7]. In addition in our textbooks, no standard methods to check for representativity and data interpretation are recommended, and a clear definition is missing frequently [8]. One definition of representativity is that the results of a representative sample can apply to the total population [9]. However, there is no information which characteristics have to be fulfilled for a sample to be a representative one. Even in textbooks of nutritional epidemiology it is only said: ‘The larger the sample size, the more likely it is to be representative of the population’ [10]. All authors agreed that the representativity is not achieved if the study population is biased. Biases are well defined and specified [11]. Therefore an approach to control for representativity is to control for possible biases of the study population. The kind of bias is dependent on the type of the study; e.g. the recall bias is the main problem in retrospective studies while the response bias mainly affects prospective cohort studies [10].

In this paper the representativity will be checked using the data of the first study population of the Kiel Obesity Prevention Study (KOPS). We have used cross-sectional data; thus the response bias (selection bias and drop out) is the most important bias to control for. There are two possibilities to con-
Subjects and Methods

Study Design

KOPS was started in 1996 in Kiel, Germany, and is planned to run until the year 2009. The aims of KOPS are i) to characterize the prevalence and incidence of overweight and obesity in children and adolescents, ii) to describe determinants and risk factors of childhood overweight, and iii) to assess long-term effects of ‘low level’ interventions at school for all children and in families who are at risk of obesity [12]. Therefore, KOPS is investigating children and adolescents at the age of 5–7 years, 9–11 years and 13–15 years. In this paper the cohort of 5- to 7-year-old children which were recruited between 1996 and 2001 is analyzed.

Subjects

Between 1996 and 2001 we investigated about 41% of all 5- to 7-year-old children (first graders) in Kiel. The primary schools of Kiel were the setting of the examination. The enrolment of children was done in cooperation with four public child health physicians from the city of Kiel. The public child health physicians examine all school entry children every year from January to July – co-workers of KOPS accompanied the public child health physicians randomly but not every day. The latter is because of the limited personal resources of the KOPS team. Therefore KOPS recruited only a part of children from all eligible 32 primary schools. Recruitment was performed throughout the different districts of Kiel without a preferential selection of schools. The school entry examination offered a good possibility to contact the parents directly and to motivate them to take part in the study.

The main cohort (i.e. all recruited children between 1996 and 2001) consists of 4,997 5- to 7-year-old children (2,503 boys; 2,494 girls). In these children complete data to different parameters of weight status exist. Therefore, this cohort can be used to analyze all questions concerning the weight status (first aim of KOPS). To obtain information on health-related circumstances of the family, a questionnaire to socio-demographic and lifestyle factors of the family was offered to all parents of children from the main cohort (within school entry examination). They were asked to fill it out at home and sent it back per mail. 2,631 parents (of 1,301 boys and 1,330 girls) completed the questionnaire; these are the responders for the data of the main cohort (within school entry examination). They were asked to fill out a non-response analysis whereby data of non-responders may be directly compared to that of the responders. In general all variables which are included in the data analysis have to be controlled for biases. However, this is not possible because only a limited number of information can be received from the total population or from the non-responders (otherwise they would be responders). To control for biases within the responders, an analysis of missing values has to be done. Besides all these methodological problems the practical consequences and the clinical relevance of an assessment of representativity have to be discussed considering the detected biases.

In this paper different methods will be used to analyze the representativity of the data of the KOPS study. The different possibilities to control for the response bias as well as an assessment of the representativity will be discussed.

Fig. 1. Recruitment of the main cohort and the responders of KOPS.

Weight Status of the Children

Weight was measured to the nearest 0.1 kg on a calibrated balance-beam scale with subjects wearing underclothes. Height was assessed to the nearest 5 mm. Body mass index (BMI) was calculated (weight (kg) / height2 (m²)). Weight status of the children was classified according to German BMI percentiles [13]: ≤10th percentile = underweight, >10th to <90th percentile = normal weight, ≥90th to <97th percentile = overweight, ≥97th percentile = obese. All measurements were done by trained nutritionists or trained students of nutritional science. To minimize measurement errors, the accuracy of measurements of all investigators was controlled every 4 weeks. The intra-observer coefficient of variation (cv) for weight and height were 0.03% and 0.22%, respectively (for 3 repeated measurements in 42 children). The corresponding inter-observer cv were 0.04% and 0.28% for 3 different observers in 42 children.

Weight Status of the Parents

Parental BMI was calculated from self-reported heights and weights during the school entry examination of the children, and parents were classified as ‘underweight’ (BMI ≤18.5 kg/m²), ‘normal weight’ (BMI >18.5 to <25 kg/m²), ‘overweight’ (BMI ≥25 to <30 kg/m²), or ‘obese’ (BMI ≥ 30 kg/m²). Therefore, data of the weight status of the parents were available from all children of the main cohort (unless parents did not provide the information; this was the case for 186 mothers and 626 fathers).
Assessment of Representativity of a Study Population

Socio-Economic Status (SES)

Socio-economic characterization of the children was done according to the district of Kiel in which the children were assigned to school. The KOPS cohorts could be biased because city districts differ in SES. SES of districts was defined by rate of unemployment, percentage of non-German citizens, and percentage of households receiving social benefits. For each characteristic, quartiles were calculated on the basis of data for all districts of Kiel. We defined a ‘low’ SES when at least one variable was in the upper quartile, in the case that all three rates were in the middle quartiles SES was ‘middle’. A ‘high’ SES was defined when at least one rate was in the lowest quartile. In a previous work we have shown that the classification of the SES according to the district as well as to the highest parental education was correlated and showed a high agreement [14].

Methods to Control for Response Bias

Comparison with the Total Population

Deviations of the main cohort from the total population could be shown. In this analysis data of age, sex, weight, height, and the district of Kiel in which the examination was done were used.

Non-Response Analysis

To analyze if there were differences between responders and non-responders, a non-response analysis was accomplished. Non-responders were defined as children who were recruited in KOPS (i.e. member of the main cohort) but did not complete the questionnaire of socio-demographic and lifestyle factors of the family. Weight status of children and parents as well as SES (according to districts) were used in this analysis.

Analysis of Missing Values

An analysis of missing values was accomplished for the responders and non-responders (e.g. the association of childhood overweight and nutrition would be biased if parents of overweight and obese children more frequently did not provide information on nutrition when compared with parents of normal-weight children). Data to the weight status of the main cohort were complete; therefore, an analysis of missing values was not necessary.

Statistical Analysis

The statistical analysis was performed with SPSS 10.0 for Windows (SPSS Inc., Chicago, IL, USA). Results were presented as median and interquartile range (IQR). For the analysis of biases according to the total population 95%-confidence intervals (95%-CI) were calculated to the mean values of the KOPS main cohort. This had to be done because the main cohort was included in the data of the total population. If the 95%-CI of the main cohort did not include the values of the total population, the values between the main cohort and the total population were significantly different. Within the non-response analysis, the non-parametric Mann-Whitney U test was used to determine differences between responders and non-responders. The chi-square test compared the prevalence of categorical variables. Level of significance was set at p < 0.05 (two-sided).

Results

Analysis of Response Bias within the Main Cohort

Table 1 shows anthropometric data and the distribution of weight status and SES of the total population as well as of the main cohort. Compared to the total population, the number of obese boys was higher in the main cohort. Girls of the main cohort more rarely belong to high SES families when compared to the total population.

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th></th>
<th>Girls</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total population</td>
<td>main cohort</td>
<td>total population</td>
<td>main cohort</td>
</tr>
<tr>
<td>Number of subjects</td>
<td>6,268</td>
<td>2,503</td>
<td>5,986</td>
<td>2,494</td>
</tr>
<tr>
<td>Age, years</td>
<td>6.2</td>
<td>6.3* (6.3–6.3)</td>
<td>6.2</td>
<td>6.2 (6.2–6.2)</td>
</tr>
<tr>
<td>Weight, kg</td>
<td>23.1</td>
<td>23.3 (23.1–23.5)</td>
<td>22.7</td>
<td>22.8 (22.6–23.0)</td>
</tr>
<tr>
<td>Height, m</td>
<td>1.20</td>
<td>1.21* (1.21–1.21)</td>
<td>1.19</td>
<td>1.19 (1.19–1.19)</td>
</tr>
<tr>
<td>BMI, kg/m²</td>
<td>15.9</td>
<td>16.0 (15.9–16.1)</td>
<td>15.9</td>
<td>15.9 (15.8–16.0)</td>
</tr>
</tbody>
</table>

*According to German BMI percentiles (cut-offs: 97th, 90th and 10th percentiles; [12]).

SES*, %

- Low: 46.6 (45.9–49.9) vs. 44.8 (43.6–47.6), p < 0.05
- Middle: 20.5 (18.8–22.0) vs. 20.4 (19.4–22.6), p = 0.05
- High: 32.9 (31.6–33.4) vs. 34.8 (33.5–35.1), p < 0.05

*Significantly different from total population.
Analysis of Response Bias within the Responders

Non-Response Analysis
The non-response rate was 47.3%. 50% of the responders as well as of the non-responders were boys. Non-responders were older and heavier and had a higher BMI than responders (table 2). Non-responders were more frequently obese and overweight when compared with responders. They belonged more often to low SES families and came less frequently from families with high SES when compared with responders (table 2). The parents of the non-responders were younger and more often overweight and obese when compared with parents of responders (table 2).

Analysis of Missing Values
The analysis of missing values within the responders (table 3) shows that the rate of missing values differed between the variables from 0.7% (weight status of mothers) to 11.4% (smoking habits of parents). There were no biases according to the weight status of the children, i.e. parents of obese and overweight children did not skip more frequently a question in the questionnaire when compared to parents of normal weight children.

Discussion
The main results of our study are: Significant differences between study participants and non-participants were detected in the main cohort as well as in the responders of KOPS. However, the direction of differences differed: while in the main cohort the prevalence of obese children and the prevalence of children from high SES families was higher than in the total population, fewer responders were overweight and obese and more frequently belonged to high SES families.

Assessment of the Representativity of the KOPS Cohorts
The consequences of an assessment of representativity have to be discussed considering the detected biases. Some authors distinguish between an ‘accurate’ and a ‘relative’ representativity.
'Accurate' means that the variables of the study population do not differ from those of the total population. A 'relative' representativity allows some differences but the degree of difference is not defined [17]. To assess the representativity of a study population, it is important to reflect the epidemiological relevance of the detected differences. The epidemiological relevance of between-group differences cannot only be deduced from statistical significant differences because in large studies already small differences often reach statistical significance. Therefore, the epidemiological relevance has to be interpreted depending on the difference that is relevant for the hypothesis of the study [18].

Taking theory into practice, we will now discuss the representativity of the KOPS cohorts.

Main Cohort

The comparison with the total population detected only differences in the prevalence of obese boys and girls from high SES families between the main cohort and the total population (table 1). There were no differences in median of BMI between the total population and the children of the main cohort (table 1). This is contrary to other studies which had systematic errors when the participation was voluntary. In the German National Health Interview and Examination Survey 1998 the body weight was higher in female non-responders when compared with responding females. In addition, participants from low social class rarely participated in the study [19]. A possible reason for our non-biased main cohort is that about 78% of the non-participants were children which were not contacted (ineligible) and therefore were not supposed to be biased. Our high participation rate of contacted children (about 75%) possibly resulted from the direct personal contact to the parents. Altogether there were very few differences between the main cohort and the total population; although there were statistical significant differences in the prevalence of obese boys and girls from high SES families, the differences were so small that they were not epidemiologically relevant. The main cohort is therefore suggested to be representative for the total population of 5- to 7-year-old children in Kiel.

Responders

The analysis of non-response showed that the responders differed systematically from the main cohort with regard to weight status of children and parents as well as SES (table 2). The detected biases have been also seen in other cross-sectional studies. Shepherd et al. [7] analyzed that main factors differentiating between non-responders and responders were social class and smoking habits. In the study of Brussaard et al. [5] the educational level among participants was higher than among non-responders. The lower participation rate of overweight and obese people is already discussed above. We additionally observed that not only the weight status of the children themselves but also that of their parents influenced the participation rate: parents of non-responders were more frequently overweight and obese than parents of responders (table 2). To detect biases within the responders, an analysis of missing values was performed. No further biases could be observed (table 3).

Taken together the responders showed biases according to SES and weight status. Therefore the data show at best a 'relative' representativity for all 5- to 7-year-old children in Kiel.

### Table 3. Number of missing values in outcome variables of responders of KOPS stratified by weight status of the children (prevalence)a

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of subjects</th>
<th>% (n) stratified by weight status of the children</th>
<th>% (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>总数</td>
<td>正常</td>
<td>超重</td>
</tr>
<tr>
<td>Number of subjects</td>
<td>2,631</td>
<td>106</td>
<td>161</td>
</tr>
<tr>
<td>Weight status mother</td>
<td>0.7 (18)</td>
<td>1.9 (2)</td>
<td>1.9 (3)</td>
</tr>
<tr>
<td>Weight status father</td>
<td>10.0 (262)</td>
<td>13.2 (14)</td>
<td>10.0 (16)</td>
</tr>
<tr>
<td>Smoking habits parents</td>
<td>11.4 (301)</td>
<td>17.0 (18)</td>
<td>13.7 (22)</td>
</tr>
<tr>
<td>SESquarters</td>
<td>11.1 (29)</td>
<td>0.9 (1)</td>
<td>0.6 (1)</td>
</tr>
<tr>
<td>SESeducation</td>
<td>2.5 (66)</td>
<td>5.7 (6)</td>
<td>0.2 (4)</td>
</tr>
<tr>
<td>Birth weight</td>
<td>10.6 (279)</td>
<td>10.4 (11)</td>
<td>8.1 (13)</td>
</tr>
<tr>
<td>Breastfeeding</td>
<td>3.8 (99)</td>
<td>4.7 (5)</td>
<td>1.2 (2)</td>
</tr>
<tr>
<td>Activity</td>
<td>1.0 (25)</td>
<td>0 (0)</td>
<td>0.6 (1)</td>
</tr>
<tr>
<td>Inactivity</td>
<td>1.5 (40)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Nutrition</td>
<td>8.1 (212)</td>
<td>8.5 (9)</td>
<td>13.0 (21)</td>
</tr>
</tbody>
</table>

*aChi-square test for significant differences between weight status (p > 0.05 for all variables)
Comparison of Methods Used

In this paper different methods to prove for biases in outcome variables were implemented in the study populations of KOPS. In general only one method will be accomplished. A comparison with the total population is useful to analyze representativeness of the main cohort. However, the subcohort of responders has to prove separately because this is the dataset with which data analysis will be done. Indeed, as shown for the KOPS populations, the main cohort is representative while the responders showed biases and could therefore called at best ‘relatively’ representative.

A comparison between the two methods used would lead to the result that the analysis of non-response is the most sensitive method to detect differences between participants and non-participants. This can be explained by the fact that in the non-response analysis two groups were compared which had no intersection (i.e. no child belonged to both groups). In contrast, in the comparison with the total population the study population itself was included in the total population so that the between-group differences were smaller than in the non-response analysis. These discrepancies between methods lead to another problem: the two groups (study population and non-participants) were included.

In this paper different methods to prove for biases in outcome variables were implemented in the study populations of KOPS. In general only one method will be accomplished. A comparison with the total population is useful to analyze representativeness of the main cohort. However, the subcohort of responders has to prove separately because this is the dataset with which data analysis will be done. Indeed, as shown for the KOPS populations, the main cohort is representative while the responders showed biases and could therefore called at best ‘relatively’ representative.

Acknowledgements

This work was supported by grants from Deutsche Forschungsgemeinschaft, Karlsruhe, and WCRF.

Disclosure

The authors declared no conflict of interest.

References


