

Estimating levels and trends in alcohol use – investigating the validity of estimates based on Norwegian population surveys

Ståle Østhus and Ellen J. Amundsen

Norwegian Institute for Drug and Alcohol Research, Oslo, Norway

Correspondence: Ståle Østhus, Norwegian Institute for Alcohol and Drug Research/SIRUS, P.O. Box 565 Sentrum, NO-0105 Oslo, Norway
E-mail: so@sirus.no Telephone +47 22 34 02 58/ Mobile +47 402 23 094 Telefax +47 22 34 04 01

ABSTRACT

Estimates of alcohol use from a series of cross-sectional face-to-face surveys, conducted by Synovate Norway on behalf of the Norwegian institute for alcohol and drug research during the 1990s and 2000s (the Substance Use Surveys, SUS), are compared with registered sales statistics of alcohol and estimates of alcohol use from Statistics Norway's Health Surveys (HS). The results show that SUS estimates of levels and trends in alcohol use are in conflict with these alternative data sources, also when standard adjustment strategies (using post-stratification weights, controlling for background characteristics in regressions) are used. We conclude that there is likely selection on alcohol use and other factors into the SUS samples, to a higher degree than in the HS samples, which renders standard estimates of alcohol use from SUS data unreliable. In fields such as substance use research, it is notoriously difficult to measure the phenomena we are interested in, and it is especially important to assess the validity of the survey estimates with data from alternative sources.

INTRODUCTION

Population surveys provide a rich and widely used data source on topics related to the use of alcohol and illegal drugs, and their links to policy. The main idea is usually to select a representative sample of the target population and make measurements yielding a "mini" picture which can be generalized to the population level. However, the road from survey estimates to population estimates is usually not straightforward. There are many types of errors present in such surveys which have to be recognized, often divided into two main categories (Groves et al. 2004). The first category is errors related to the persons and includes coverage error (the sampling frame is not identical to the target population), sampling error (a sample is studied, not all persons in the frame), non-response error (some persons do not answer) and adjustment error (e.g. weighing for under-represented groups to repair weaknesses). The second category is errors related to measurements and includes validity (that the measures are valid for the underlying construct), reliability (differences between the true value for the person and the response recorded) and processing error (the responses are edited, mainly to repair weaknesses).

It is recognized that in the process of measuring substance use with surveys, some types of errors are more important than others (Del Boca & Darkes 2003; Gmel & Rehm 2004). Heavy users may be marginalized, without a permanent address or a registered telephone, and therefore not included in a sampling frame of an official housing or telephone register. Or they may be more difficult to reach, for example due to a higher propensity to be in treatment or in jail, which also

makes non-response error important. Measurement errors also occur. Questions on the use of illegal drugs, and heavy drinking, are sensitive. Some groups may report a higher level than the true one and some report a lower level, but for the population at large it is reasonable to assume that actual consumption is underreported in most surveys. For alcohol the coverage of the total consumption in most surveys is said to be 40 to 60 percent, although significant improvements in coverage can probably be achieved with better measurement strategies (Greenfield & Kerr 2008).

Because the actual consumption of alcohol and illegal drugs in the population is not known there is no "gold standard" to which survey estimates can be compared. A feasible, albeit imperfect, strategy is then to compare survey estimates with estimates based on alternative data sources that can be used to indicate levels and trends in actual consumption. For alcohol, but not for illegal drug use, one obvious such alternative data source is statistics on registered sales. One advantage of these data is that we can be almost certain that actual consumption is higher than registered sales, e.g. due to unregistered consumption of imported alcohol (Nordlund 2000; 2003). It is also reasonable to assume that unregistered consumption of liquor and wine is substantially higher than unregistered consumption of beer. For example, the estimated privately imported/registered sales proportions are less than 5% for beer, around 15% for wine and more than 30% for liquor and did increase for all beverage types during the 1990s according to different surveys (cf. Nordlund 2003). Thus, we have a comparison standard for survey estimates of alcohol consumption that is not perfect, but where some of the errors are more or less known.

Another alternative data source to which survey estimates can be compared is other surveys on the same topic. Here, we focus on a long time-series of cross-sectional face to face surveys on the use of alcohol and illegal drugs (the substance use surveys, SUS) conducted by the Norwegian Institute for Alcohol and Drug Research (SIRUS) since 1956, the last ones carried out in 1994, 1999, 2004 and 2009. The surveys have been an important source of information on trends and levels in substance use, and attitudes towards alcohol and drug policy, in Norway (e.g. Horverak & Bye 2007; Østhus 2005). In an earlier study based on surveys conducted in 1978 and 1991 it was established that alcohol use estimates were fairly similar to those obtained from available alternative data sources (Nordlund 1992). However, in recent years, there has been growing concern over the ability of these surveys to produce valid population estimates (of alcohol consumption in particular) in part because trends in survey estimates were in conflict with trends in registered sales. We compare SUS estimates to estimates from the Norwegian Health Surveys (HS), which are large population surveys conducted by Statistics Norway during approximately the same period as the four SUS (in 1995, 1998, 2005 and 2008)¹. Information on substance use in these data are from a postal questionnaire where questions on alcohol and illegal drug use are asked in addition to a set of sensitive health questions. Advantages of these data include the fact that they are well documented (see <http://www.ssb.no/emner/00/90/levekaar/> for documentation in Norwegian), with information on sampling strategies as well as indicators of representativeness. Differences between the gross and the net sample with respect to socio-demographic characteristics (gender, age and region) are also fairly small in these surveys.

Although alcohol use is measured in both sets of surveys, there are differences in measurements between them, with the SUS having much more nuanced questions on alcohol use (e.g. questions on beverage-specific consumption) than the HS. With the availability of series of surveys over time with different sampling frames, wording and design of questions about substance use, the aim of this study was to compare results regarding alcohol use and discuss the strengths and weaknesses of each series. Our main hypothesis is that these surveys, to varying degrees, failed to provide samples that were representative of the population, rendering estimates of substance use (e.g. means and proportion) that cannot be validly generalized to the target population. Thus, it is errors related to the sampling of persons more than measurement errors that are our main concern.

SAMPLING IN THE SUBSTANCE USE SURVEYS (SUS) AND THE NATIONAL HEALTH SURVEYS (HS)

A key difference between SUS and HS has been their sampling strategies. The SUS respondents were selected through a three-step procedure: a master sample of

municipalities was first selected following a stratification of all Norwegian municipalities into 17 strata by region, number of inhabitants and main source of employment. In each municipality selected, a random selection of start addresses was drawn from land line telephone registers. Finally, from each start address, the interviewers were instructed to go to four new addresses following a specific system, trying to recruit the household member with the most recent day of birth, and who were over the age of 15 and at home when contacted. No interviews were supposed to be carried out at the start addresses. In principle, this method will result in a representative sample of the population aged 16 and over. However, according to a documentation report from Synovate Norway on the 2009 SUS data (available upon request from the authors), the interviewers were successful in recruiting respondents at less than 20% of the addresses they visited. This suggests that non-response is a serious problem in these surveys. No documentation is available for the earlier SUS surveys. In addition, there is a general decline in land line telephone subscriptions in Norway, as there is in many other countries. Thus the sampling frame cannot be assumed strictly representative for the population.

This sampling strategy leaves much of the responsibility of recruiting respondents in the hands of interviewers in the field. Survey participation is conditional on whether potential respondents are actually at home and willing to participate when interviewers visit their neighborhood. This is unlikely to be random, and likely to be related to potential respondents' drinking pattern. For example, the number of completed interviews are unevenly distributed across the days of the week, with most interviews (around 70%) conducted Monday to Thursday in all four surveys². An even distribution over weekdays implicates around 56% to be interviewed on these days. In addition, most interviews were conducted after four pm, and this varied between the surveys over time. The proportion of interviews that were conducted before four pm on a weekday (i.e. during standard working hours) were around ten percent in 1994 and 1999, increasing to 26% and 20% in 2004 and 2009 respectively. During this time of day and week, interviewers are more likely to encounter particular social groups that are at home and willing to participate, and who may have deviant drinking or illegal substance use patterns. Examples are unemployed or parents of small children.

More generally, because people's drinking pattern typically varies with time and day of the week, survey participation is likely to be related to drinking levels. It is for example plausible that the more often people drink alcohol, the less likely they are to be willing to participate (or at home) when interviewers arrive at their door. Although much of the drinking in Norway takes place during weekends and evenings, people who drink frequently also drink during weekdays. Therefore, the sampled respondents are a non-randomly

selected group of people that tend to be home and willing to participate during weekdays and evenings. Selection on substance use behaviors into the sample, however, can be both positive (e.g. oversampling of heavy users) and negative (e.g. oversampling of abstainers). It is therefore difficult to determine a priori how this selection has an impact on estimates of means or proportions. It seems likely, however, that this selection is dependent on aggregate consumption levels. When aggregate consumption is high, the SUS estimates of drinking levels will tend to be more downwardly biased because the sampling procedure favors nonparticipation of people who are drinking more frequently. Thus, the SUS samples will tend to be less representative of the population of alcohol users the more alcohol users there are, and the more people drink.

Another likely source of selection into the SUS samples is regarding socio-economic status, which we know is related to drinking habits and other substance use (Babor et al. 2010, Huckle et al. 2010). Interviewers are instructed to contact at most four addresses before the sampling starts over from a new initial address in another area, thus ensuring that all households have the same probability of being included in the sample. Interviewers are paid per interview, rather than per hour of fieldwork, which implies that there are incentives to reduce non-response. This also means that there are incentives to deviate from a complex, time-consuming sampling pattern. Rather than repeatedly starting over from new start addresses, a more effective strategy for the interviewer would be to contact as many households as possible in the same residential area in order to fill the needed quota of interviews. These incentives are likely stronger in low-income residential areas, where population density tends to be higher (e.g. people are more often living in apartment buildings than in detached houses), and where participation rates tends to be smaller. Thus households with lower level socio-economic status may be sampled too often. There is no documentation that interviewers violate their instructions, however, so the description is merely of the systems incentives, not actual behavior of interviewers. The HS all use random samples drawn from population registers, where all Norwegian citizens are listed. Thus, errors related to the sampling frame or sampling procedures are not likely to be a large problem in these surveys. Not everybody is living at their listed address, however, and especially marginalized persons with heavy substance use may stay in shelters, rehabilitation premises or be in treatment for long periods. An information letter and brochure about the survey and the series of Living Condition/Health Surveys was sent to everyone in the gross sample, and reminder letters were also sent off. The HS includes face-to-face or telephone interviews with respondents in addition to the postal questionnaire that were sent to everyone in the gross sample, so the interview burden can be substantial. Economic incentives to participate are given in the form of a lottery of gift

certificates (e.g., in 2008, two NOK 10 000 gift certificates and ten NOK 1000 gift certificates) while moral incentives are given in statements about the importance of the survey and that replacing the sampled individual with another person will produce biased estimates. The response rate for the questionnaire on alcohol and illegal drug use has declined over time, from 86% in 1995³ and 72% in 1998 to 57% and 50% in 2005 and 2008 respectively, creating increasing problems of non-response.

Note that there is little reason to suspect participation in the HS to be directly related to alcohol or illegal drug use, as was argued above for the SUS. There is, however, reason to suspect that HS participation is related to factors that are linked to substance use, such as socio-economic status, having a permanent address or language skills. For example, the tendency of the highly educated to be more willing to participate in surveys is also relevant for the HS. Due to the falling response rate we may also have successive cross-sectional samples that are increasingly less representative of the population at large. However, according to the documentation of the HS data there are only small differences between the gross and the net sample in characteristics such as gender, age or region. All things considered, it seems reasonable to assume that the HS samples on average are *more* (but not perfectly) representative for the Norwegian population than the SUS samples.

In both SUS and HS, heavy drinkers are likely to be underrepresented compared to the target population (Tolonen et al. 2010). One reason is that heavy drinking may influence the probability of being home at weekday evenings (SUS). Another reason is due to heavy drinking reducing the willingness to participate in both face-to-face and postal surveys, e.g. due to impaired mental or physical health and wellbeing. By contrast, heavy drinkers are likely to drive up registered sales statistics. This is because the distribution of alcohol consumption is highly skewed, with a small proportion of drinkers being responsible for a large proportion of the total consumption (Skog 1985). In Norway, it is estimated that around ten percent is responsible for around half of the total consumption, although this ratio is neither fixed over time nor across different social groups (ibid). The skewed distribution of alcohol consumption in the population, together with the usual undersampling of heavy drinkers in survey samples, are important reasons why survey estimates of alcohol use are typically much lower than what is indicated by sales statistics.

When means or proportions are estimated from the SUS, it is standard to use post-stratifying weights (calculated by Synovate Norway) to adjust for differences between the sample and population distributions of observed person characteristics (gender, age and geographic region). However, it must be recognized that using such weights are no panacea against sampling bias. In particular, using such weights will not produce valid

population estimates if there is selection into the sample on factors that are not included in the calculation of the weights (e.g. alcohol use). In fact, adjusting for a selection of observed factors such as age or gender may produce a sample that is less representative with respect to characteristics that are not included in the weights, and they are often useless when selection is on unobserved characteristics.

MEASURES OF ALCOHOL USE

Within each set of surveys, the measures of alcohol use are virtually identical across time. By collapsing response categories in the respective surveys, similar measures can also be constructed across the survey series. However, there are real differences in the way questions and response categories are formulated in the two sets of surveys. For example, in the HS, the respondents are asked to report how often he or she "have been drinking alcohol during the past 12 months" by using six response categories ranging from "never" to "4-7 times per week"⁴. In the SUS, the respondents are first asked if he or she "have tasted beer" during the past 12 months, and then asked to report how many times per week, month or year he or she have tasted beer. The same procedure is repeated for wine and liquor (and in the last two surveys also for alcopops), resulting in multiple continuous measures of drinking frequency (one each for beer, wine, liquor and alcopops). It is possible that the more nuanced beverage-specific questions in the SUS will drive up frequency estimates compared to the coarser HS question⁵. However, it is also possible that the response categories used in the HS will drive up reported frequency. The three highest response options all indicate very high frequency use (times per week) and may therefore "normalize" such use and thus reduce respondents' unwillingness to report actual high-frequent use that may otherwise have been deemed socially undesirable (Greenfield & Kerr 2008).

It should also be noted that the SUS measure of annual alcohol consumption, which we use in our comparison with registered sales data, has some well-known disadvantages. Annual alcohol consumption is measured with the "frequency-quantity" (QF) method. In the QF method the reported beverage-specific consumption frequency is multiplied by the reported "typical amount drunk". The latter is calculated from a nuanced categorical measure of beverage-specific standard drinks. Information on quantity is important for understanding drinking patterns, and the SUS data is therefore in principle better suited to describe alcohol use than surveys that rely on one-dimensional frequency measures. However, it has long been recognized that the reported usual quantity tends not to be the arithmetic mean of a person's varying drinking pattern; It typically under-represents heavy drinking occasions and leads to downward bias in estimated volume (Greenfield & Kerr 2008). Thus, the measuring strategy

is also likely to be responsible for differences between SUS estimates and registered sales data.

The questions on how often the respondent have been drinking to intoxication during the past 12 months are very similar in the two sets of surveys, but again the response categories differ. In the HS, the same six response categories are used as in the question on drinking frequency. In the SUS, the respondent is asked to report intoxication frequency as the number of occasions during the past 12 months, yielding a continuous measure. Because social desirability may be important for respondents when answering this question (in particular in the context of a face-to-face interview such as SUS), the open-ended continuous SUS measure may underestimate respondents' actual intoxication frequency. By contrast, the response categories used in the HS may both normalize high-frequent intoxication and the fact that the HS are a postal survey may remove much of the social desirability effect.

EMPIRICAL STRATEGY

In the present study, simple means and proportions based on the SUS data are compared with data on registered sales and alcohol use estimates from the HS data. Characteristics of the SUS and HS samples are given in table 1. Annual consumption is calculated with the QF method, following a standard procedure used by researchers at the Norwegian institute for Alcohol and Drug research when results have been presented from these data (Horverak & Bye 2007). To obtain estimates of the proportion of last year abstainers and monthly drinkers from the SUS data, we first calculated respondents' drinking frequency as the number of occasions he or she had tasted either beer, wine or liquor during the past 12 months⁶. Abstention is defined as last year abstention in both sets of surveys (not having tasted beer, wine, liquor or alcopops during the past 12 months in the SUS). Monthly drinking is similarly defined as having used alcohol more than once per month during the past 12 months, in the HS measured by collapsing response categories "2-3 times

Table 1. Key figures for surveys.

	Substance use surveys (SUS)			
	1994	1999	2004	2009
Contact	–	–	–	11340
Net sample	2752	2170	3191	2073
	–	–	–	18,3%
	Health surveys (HS)			
	1995 ¹	1998	2005	2008
Gross sample	7330	9735	9187	9684
Net sample	6263	7004	5212	4832
	85.4%	71.9%	56.7%	49.9%

¹ The 1995 HS sample was a household sample (gross/net sample = 5880/4422 households). A postal questionnaire on alcohol use was sent to all sampled persons 14-79 years old (N=7330).

per month” and higher, and in the SUS measured as having tasted beer, wine, liquor or alcopops 13 times per year or more often⁷. Intoxication frequency is measured from questions on how often respondents have been “clearly intoxicated” from alcohol use (measured with a single question in both surveys) during the past 12 months; 2-3 times per month or more often in the HS and 13 times per year or more often in the SUS. Abstainers are not included when the proportion of monthly intoxicated drinkers are calculated.

Post-stratifying weights are used when convention dictates it (i.e. in the calculation of means and proportions from both SUS and HS). Because the first HS survey (from 1995) is a household sample, household weights are used to ensure that the sample is representative for the population at the individual level. All the weights we use in this study are weights that are already calculated, and included in the data, by either Synovate or Statistics Norway. No weights are included in the 2008 HS data, so all estimates based on these data are unweighted estimates. In addition to measures of alcohol use, we also compare SUS sample distributions of other characteristics (factors that are not included in the calculation of the post-stratifying SUS weights) with population estimates from other data sources: Employment status in the SUS (measured as having part or full time employment at the time of the interview) is compared with employment status in the Norwegian Labor Force surveys (measured as having paid work in the survey week). Marital status and educational attainment in the SUS are compared with the relevant population distributions of marital status and educational attainment taken from population registers. Documentation of these data can be found at www.ssb.no. When SUS and HS estimates of drinking frequency and intoxication are compared, adjusted predictions from logistic regressions with a set of background characteristics (gender, age, educational attainment and geographic region) entered as independent variables are presented alongside unadjusted estimates. This is done in order to investigate whether differences between the two sets of surveys disappear after control for observed characteristics that are plausibly linked to survey participation as well as alcohol use. Stata 11 was used for all analyses. Adjusted predictions (or predictive margins) with appropriate standard errors were obtained with the `-margins` routine (StataCorp 2009).

RESULTS

Table 2 shows estimated per capita annual consumption of pure ethanol in the SUS and registered per capita annual sales in pure ethanol for corresponding years by beverage type. The survey estimates of per capita consumption is lower than per capita sales when total consumption is compared to total sales as well as when beverage-specific consumption is compared with beverage-specific sales. Consistent with the likely lar-

ger share of unregistered consumption of liquor than of beer (Nordlund 2000), differences between sales and self-reported consumption are on average larger for beer than for liquor. One exception is the 2009 survey, where the survey estimate of liquor consumption is unusually low (54% of sales) compared to previous surveys (ranging from 84% to 97% of sales). Because the sharp increase in wine sales is not reflected in the survey estimates of wine consumption, the wine consumption/sales ratio is declining over time (from 73% in 1994 to 35% in 2009).

Trends in self-reported consumption also differs substantially from trends in registered sales, both in terms of overall trends and changes between survey years. Registered sales have grown, albeit somewhat unevenly, over the period. Per capita total sales of beer, wine and liquor increased from 4.7 to 6.6 liters of pure ethanol between 1994 and 2009 (an increase of 38%), with wine sales increasing much more (130% overall) than sales of beer (8%) or liquor (31%). Trends in self-reported consumption are inconsistent with trends in sales for total consumption as well as beverage-specific consumption. While registered total and beverage-specific sales have increased overall, and have increased or remained stable between all survey years, self-reported consumption sometimes increase and sometimes decrease. The sharp increase in wine sales is not fully reflected in the survey data.

Table 2. Per capita alcohol consumption per person 16 years and over in liters pure ethanol by beverage type. Registered sales and substance use surveys estimates.

	Year	Registered sales		Self-reported consumption		
		Mean	Change	Mean	95% C.I.	Change
Total	1994	4.74		3.29	[3.1–3.5]	
	1999	5.38	13.5%	3.24	[3.0–3.5]	–1.6%
	2004	6.11	13.6%	3.70	[3.5–4.0]	14.5%
	2009	6.55	7.2%	3.12	[2.8–3.4]	–15.7%
	Overall	5.70	38.2%	3.34		–5.1%
Beer	1994	2.75		1.60	[1.5–1.7]	
	1999	2.84	3.3%	1.50	[1.3–1.7]	–6.4%
	2004	2.96	4.2%	1.73	[1.6–1.9]	15.6%
	2009	2.96	0.0%	1.61	[1.4–1.9]	–7.2%
	Overall	2.88	7.6%	1.61		0.4%
Wine	1994	0.99		0.72	[0.7–0.8]	
	1999	1.49	50.5%	0.85	[0.7–1.0]	17.8%
	2004	1.90	27.5%	0.87	[0.8–0.9]	1.9%
	2009	2.28	20.0%	0.81	[0.7–0.9]	–7.0%
	Overall	1.67	130.3%	0.81		11.7%
Liquor	1994	1.00		0.97	[0.9–1.1]	
	1999	1.05	5.0%	0.89	[0.8–1.0]	–8.2%
	2004	1.25	19.0%	1.10	[1.0–1.2]	24.7%
	2009	1.31	4.8%	0.71	[0.6–0.8]	–35.9%
	Overall	1.15	31.0%	0.92		–26.7%

Notes: Registered sales/reported consumption of alcopops is not included. Change is the percentage growth in sales/consumption relative to the previous measuring year.

Source: AS Vinmonopolet and own calculations based on (weighted) SUS data.

Table 3. Drinking frequency and intoxication frequency in two sets of population surveys.

	Year	Last year abstainer			Drink monthly			Drink weekly			Intoxicated monthly		
		HS	SUS	Diff.	HS	SUS	Diff.	HS	SUS	Diff.	HS	SUS	Diff.
Unadjusted	1994/1995	16.6	12.8	-3.8 ***	48.9	55.6	6.7 ***	30.8	29.8	-1.0	24.2	9.7	-14.5 ***
	1998/1999	13.9	10.0	-4.0 ***	51.7	58.7	7.0 ***	31.2	31.8	0.6	19.8	8.7	-11.1 ***
	2004/2005	11.7	11.7	0.0	59.5	58.8	-0.7	34.6	33.2	-1.4	17.5	10.7	-6.8 ***
	2008/2009	11.2	15.6	4.4 ***	63.5	56.2	-7.4 ***	38.3	34.5	-3.8 **	16.1	8.9	-7.2 ***
Adjusted predictions	1994/1995	15.9	11.3	-4.6 ***	50.0	57.1	7.1 ***	31.9	30.7	-1.2	23.4	8.5	-14.8 ***
	1998/1999	13.9	9.3	-4.6 ***	51.7	59.7	8.0 ***	31.4	32.7	1.2	19.1	8.5	-10.5 ***
	2004/2005	11.6	11.3	-0.2	58.8	59.9	1.0	34.0	34.3	0.3	18.0	10.4	-7.6 ***
	2008/2009	10.8	15.8	5.1 ***	64.0	55.3	-8.8 ***	36.9	33.3	-3.6 **	18.7	9.1	-9.6 ***

Notes: HS=Health survey, SUS=Substance use survey. Adjusted predictions from logistic regressions with gender, age, age squared, educational attainment (three cat.), geographic region (six cat.) and interactions between year and these variables entered along with a survey dummy, year dummies and interactions between survey and year entered as independent variables. F-test of differences between unadjusted proportions. Tests of differences between adjusted predictions based on delta method standard errors. All estimates from weighted data.

* $p < 0,05$; ** $p < 0,01$; *** $p < 0,001$.

Table 4. Person characteristics in official statistics and substance use surveys.

	1994			1999			2004			2009		
	OS	SUS		OS	SUS		OS	SUS		OS	SUS	
		Unweighted	Weighted		Unweighted	Weighted		Unweighted	Weighted		Unweighted	Weighted
Percent women (Ages 16-79)	50.2 (52.1)	49.4 [47.5-51.3]	50.8 [48.7-52.9]	50.1 (50.0)	52.0 [49.9-54.2]	50.4 [48.1-52.8]	49.9 (48.7)	52.0 [50.2-53.8]	50.5 [48.7-52.4]	49.5 (53.4)	51.7 [49.4-53.9]	50.2 [47.7-52.7]
Mean age (Ages 16-79)	43.1 (43.8)	41.0 [40.4-41.6]	43.3 [42.6-44.1]	43.5 (43.2)	43.0 [42.2-43.7]	43.8 [42.9-44.6]	43.9 (44.6)	42.2 [41.6-42.8]	44.1 [43.5-44.7]	44.1 (46.9)	45.0 [44.2-45.8]	44.5 [43.6-45.3]
Percent in paid employment (Ages 16-74)	64.9 (72.9)	64.3 [62.4-66.1]	61.0 [58.9-63.2]	71.0 (63.4)	68.8 [66.8-70.9]	67.8 [65.5-70.1]	69.3 (71.9)	63.1 [61.3-64.9]	63.5 [61.7-65.3]	70.5 (75.2)	59.6 [57.4-61.9]	63.3 [60.8-65.8]
Percent married (Ages 20-79)	55.1 (58.2)	45.1 [43.2-47.0]	47.5 [45.4-49.7]	52.6 (55.9)	48.7 [46.4-50.9]	50.1 [47.7-52.5]	50.8 (52.0)	41.6 [39.7-43.4]	43.4 [41.5-45.3]	48.9 (58.0)	46.2 [43.9-48.6]	46.6 [44.0-49.3]
Percent with univer- sity level education (Ages 16-66)	20.5 (26.3)	35.0 [33.1-36.9]	30.0 [28.0-32.0]	24.0 (25.1)	32.5 [30.4-34.7]	33.0 [30.6-35.4]	26.6 (27.8)	31.2 [29.5-33.0]	30.7 [28.9-32.5]	29.8 (36.8)	31.0 [28.8-33.2]	35.5 [32.8-38.1]

Notes: OS=Official statistics (National Labor Force surveys, marriage and education registers, see www.ssb.no for documentation), SUS=Substance use surveys. Weighted distribution in Health Surveys in parentheses. 95% confidence intervals of SUS estimates in brackets.

Table 3 shows unadjusted and adjusted prevalence of drinking and drinking to intoxication in samples from SUS and HS. Overall, the estimated proportions of last year abstainers, weekly or monthly drinkers, and intoxicated monthly differ significantly between the two survey samples. Similar to what is shown in table 2, there is no clear trend in the deviations between the estimated proportions from the two surveys. In the two first survey periods (1994/1995 and 1998/1999), the proportion of last year abstainers is lower and the proportion of monthly drinkers higher in the SUS than in the HS, while the reverse is true for the last survey period (2008/2009). The estimated proportion of monthly intoxication is significantly lower in the SUS than in the HS all four years, but the size of this difference declines over the period (i.e. it is larger in 1994/1995 than in 2008/2009). When the predicted proportions were adjusted for observed characteristics that may plausibly be linked to the sampling proce-

dures of the two surveys (i.e. gender, age, education and geographic region), the differences between the predictions from the two surveys are similar in direction and magnitude to the unadjusted estimates (cf. lower panel of table 3).

Table 4 show the sample gender and age distributions, together with sample distributions of characteristics (i.e. employment status, marital status and educational attainment) that are not accounted for by the post-stratifying weights included in the SUS data, and the corresponding population distributions according to alternative data sources from Statistics Norway (with HS estimates in parentheses). The sample distributions differ from the population distributions of all three characteristics that are not included in the weights, suggesting that the SUS samples are not representative of the population in these respects. The problem seems especially pronounced with respect to educational attainment, where all the SUS samples includes a higher

proportion of persons with university level education than the population distributions according to official statistics (in this case, register data on the educational attainment of the entire population, www.ssb.no). Over-sampling of highly educated persons is also evident in the HS samples, however, especially in the first (1995) and last (2008) of these surveys. Thus, neither set of surveys can be claimed to accurately represent the population distribution of educational attainment.

The SUS also includes fewer in paid employment than reported in the National Labor Force surveys (NLF)⁸, with larger differences between survey and official statistics in the last two surveys (in 2004 and 2009). In 1994 and 1999, the SUS proportions in paid employment are similar to the population proportions as estimated by NLF. However, consistent with the larger share of interviews during standard work hours in the 2004 and 2009 surveys, the proportion in paid employment is downwardly biased in these later samples. The HS samples also differ from the NLF samples all years, but no clear pattern is evident for these deviations. Compared to marital status registers, the SUS samples also include more unmarried persons, with larger differences between survey and official statistics in 1994 and 2004. Married persons tend to be over-represented in all the HS samples. Note that differences between the sample distributions and the population distributions are not always smaller in the weighted than in the unweighted survey samples. In the last survey, for example, the weighted proportion of highly educated persons is even higher than the unweighted proportion, resulting in a greater difference between population distributions and the weighted sample than between population distributions and the unweighted sample.

DISCUSSION

In this paper we compared different measures of alcohol consumption in a series of cross-sectional surveys on substance use (SUS) conducted by Synovate Norway on behalf of the Norwegian Institute for Alcohol and Drug Research during the 1990s and 2000s to data on registered sales and data from Statistics Norway's health surveys (HS). There were clear differences in both levels and trends in consumption when SUS estimates were compared to these alternative data sources. In addition, sample distributions of characteristics that are linked to alcohol consumption in the SUS were compared to population distributions of these characteristics as given by survey and register data produced by Statistics Norway. The samples in the SUS are not representative for the population with respect to these characteristics, and the post-stratifying weights do not fully account for biases. The most likely conclusion from our investigation, is that there is selection on alcohol consumption and other, possibly unobserved, characteristics linked to alcohol consumption into the SUS samples that will bias estimates that are typically

reported based on these data.

One can argue that neither statistics on registered sales nor estimates from the health surveys offer reliable data on actual consumption in the population. Perhaps we cannot trust *any* estimates and, consequently, that knowledge of population trends and levels in substance use is simply out of reach for statistical analysis. This is not our view. Granted, alcohol consumption is expected to be higher than registered sales because of "unregistered consumption" of imported or unlicensed alcohol that have to be added to sales. Trends in registered sales may also reflect shifts between registered and unregistered consumption, suggesting that survey data – in theory at least – can provide better estimates of actual consumption. The SUS data, however, show inconsistent patterns in trends in self-reported alcohol consumption – and differs substantially from trends in registered sales. This is the case also for beer, where shifts between registered and unregistered consumption is less relevant because unregistered consumption is smaller (Nordlund 2000). The sharp increase in wine sales is neither fully reflected in the SUS data. This increase seems unlikely to merely stem from such consumer shifts, in particular because both theory and empirical evidence suggests that unregistered consumption of wine has also increased (*ibid*).

One possible explanation for the missing upward trend in alcohol consumption in the SUS data, and wine consumption in particular, is that SUS participation in itself is dependent on drinking levels. As aggregate consumption increases, the SUS sampling procedure is likely to exclude a growing proportion of high-frequency drinkers from the sample, resulting in increasing downward bias in drinking level estimates. If so, it is not surprising that the missing upward trend in the SUS is especially pronounced for wine consumption, because it is here we expect the sharpest increase in consumption during the period of investigation. However, we see the same trend – or lack thereof – when SUS estimates are compared to HS estimates of drinking frequency. The HS results suggest a decline in the proportion of last year abstainers and a fairly sharp increase in the proportion of monthly and weekly drinkers over the period, but these upward trends are not similarly evident in the SUS data.

The validity of the estimated alcohol consumption from the HS can also be questioned. One aspect is the steady decline in response rates in these surveys, from 86% in 1995 to 50% in 2008. There are also no good reasons to expect that national health surveys in general are better able to provide unbiased estimates of substance use than other surveys. Nevertheless, the trends from HS during the 1990s and 2000s in reported drinking frequency are well in line with trends in registered sales. Therefore, the HS seems to provide a more valid, coarser, picture of trends in alcohol use during the past two decades than the SUS. In addition, the documentation of these data suggests that the

samples used are fairly representative of the target population with respect to relevant socio-demographic characteristics even without the use of adjustment weights. Based on our assessment of the sampling strategies used in the SUS and HS, it is also likely that selection into the SUS samples is more directly related to alcohol use than is the case of the HS.

The comparisons between the SUS and the HS rest on some untested assumptions. One is that the observed differences are mere artifacts of the survey design and do not reflect substantial differences in substance use in the underlying population the samples are drawn from. Because we compare estimates based on samples drawn for different years it is possible that differences between the surveys reflect substantial differences between those years: the SUS survey from 1994 is compared with the HS from 1995, the SUS survey from 1999 is compared with the HS from 1998 and so on. Still, there are several reasons why this is not likely. First, the magnitude of the differences between the surveys speaks against it. The differences between different surveys that are one year apart are generally much larger than differences between different surveys within the same set of surveys (i.e. either SUS or HS) that are three, five or seven years apart. Second, the direction of the differences is not consistent with the succession of the SUS and HS. For example, in 1994/1995 SUS was conducted first while HS was conducted first in 1998/1999, but the differences between SUS and HS with respect to drinking frequency and intoxication is in the same directions for both periods (cf. table 2).

Another important assumption made is that differences between estimates from the two sets of surveys are due to different sampling procedures and not merely the result of different ways of measuring alcohol use. There are real differences, however, in the way questions and response categories are formulated in the two sets of surveys. Therefore, the comparisons are not made between two sets of surveys with identical questions, but with measures constructed from different questions that are assumed to measure the same underlying concept (e.g. drinking frequency). For example, we assume that collapsing response categories from the HS measure of drinking frequency measures the same underlying concept as combining data in SUS on drinking frequency. In HS an indicator variable of drinking alcohol more than once per month was based on the five response categories “2-3 times per month” through “6-7 times per week”, while in SUS it was based on the response categories of drinking beer, wine or liquor more than once per month.

Because the measurement strategy within each set of surveys have remained the same throughout the entire period one would think that differences between the surveys would be in the same direction for all comparison periods if these differences merely stem from different measurement strategies. This, however, is not the case. In the first two periods (1994/1995 and

1998/1999) a lower proportion of abstainers and a higher proportion of monthly drinkers are reported in the SUS surveys compared with the HS surveys. There is no large differences in drinking frequency in the third comparison period (2004/2005), and in the last comparison period (2008/2009) the situation is reversed (cf. table 3). With respect to intoxication frequency, the SUS estimates are lower than the HS estimates in all four comparison periods. It seems plausible that much of this difference can be attributed to different measuring strategies. Again, however, the difference between the two sets of surveys is not constant over time, and in the SUS intoxication frequency is stable or slightly increasing whereas in the HS intoxication frequency is declining.

It is possible that the decline in intoxication frequency in the HS surveys is related to the decreasing participation rate. Intuitively (and theoretically, see e.g. Skog 1985), it does not seem very plausible that heavy drinking should become less common when drinking becomes more common. This is particularly true considering that there has been an increase in alcohol-related hospitalizations during the last decade in Norway, e.g. with hospitalization due to acute intoxication more than doubled between 1999 and 2010 (from 28 per 100.000 inhabitants 16 years or older in 1999 to 63 per 100.000 inhabitants 16 years or older in 2010). It may also be the case that respondents are more reluctant in interpreting their own drinking as problematic (e.g. leading to intoxication) in high consumption than in low consumption periods (Nordlund 2010). If so, the declining intoxication frequency in the HS may be the result of reporting bias. However, because the reverse may also be true, more research is needed on this issue.

We also compared the sample distributions of background characteristics that are plausibly linked to both alcohol use and the sampling procedure in the SUS regarding employment and marital status, and educational attainment. The results are that the distributions in the SUS samples differ from the population distributions for these characteristics in the unweighted as well as in the weighted samples. This suggests that there is a selection on these characteristics into the survey samples, and that the post-stratifying weights constructed to adjust for sampling bias does not fully account for such selection. The distributions in the HS samples, however, also differ from the population distributions of these characteristics, and it is not clear to what extent drinking level estimates from either survey series are affected by it. The survey measures of these background characteristics differed somewhat from the measures used in the other official statistics, however. For example, labor force participation in the NLF is based on questions about paid employment in the week immediately preceding the survey period while the SUS measure is based on questions on whether or not the respondent had full or part time employment at the time of the survey. In our view, these differences

should be noted, but are of minor importance.

Based on our investigation, we conclude that there is selection on alcohol use and on observed – and possibly unobserved – characteristics related to alcohol use into the SUS samples. This poses serious threats to the validity of estimates of levels and trends in alcohol use based on these surveys. One possible explanation for this is the applied sampling procedure, which may lead to oversampling of groups that are more prone to have deviant drinking patterns and undersampling of frequent drinkers. Typical adjustments made, such as applying post-stratifying weights or controlling for a range of observed background characteristics, does not seem to solve the problem. Our investigation has some

broader implications that go beyond a critique of the validity of the estimates that can be derived from this particular set of surveys. First, our investigation has demonstrated that the use of post-stratifying weights or other standard adjustments strategies (e.g. regression adjustments) does not always provide a cure for sampling bias. Second, our investigation has demonstrated the virtue of exerting caution generalizing from survey estimates to population values. In fields such as substance use research, where it is notoriously difficult to measure the phenomena we are interested in, it is especially important to assess the validity of the survey estimates with data from alternative sources.

REFERENCES

- Babor, T. F., Caetano, R., Casswell, S., Edwards, G., Giesbrecht, N., Graham, K., et al. (2010). *Alcohol: no ordinary commodity: research and public policy* (Second edn.). Oxford: Oxford University Press.
- Del Boca, F.K., Darkes, J. (2003). The validity of self-reports of alcohol consumption: state of the science and challenges for research. *Addiction*, **98**, 1-12.
- Gmel, G., & Rehm, J. (2004). Measuring alcohol consumption. *Contemporary Drug Problems*, **31**, 467-542.
- Greenfield, T.K., & Kerr, W.C. (2008). Alcohol measurement methodology in epidemiology: recent advances and opportunities. *Addiction*, **103**, 1082-1099.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeay, R. (2004). *Survey methodology*. Hoboken, N.J.: Wiley.
- Horverak, Ø. & Bye, E.K. (2007). *Det norske drikkemønsteret [The Norwegian drinking pattern]*. Oslo: Norwegian Institute for Alcohol and Drug Research.
- Huckle, T., You, R. Q., & Casswell, S. (2010). Socio-economic status predicts drinking patterns but not alcohol-related consequences independently. *Addiction*, **105** (7), 1192-1202.
- Nordlund, S. (1992). *Metoder og metodeproblemer ved estimering av alkoholforbruk [Methods and methodological problems concerning the estimation of alcohol use]*. Oslo: Norwegian Institute for Alcohol and Drug Research.
- Nordlund, S. (2000). Real consumption of alcohol: methods and errors of estimation. *Contemporary Drug Problems*, **27**, 235-251.
- Nordlund, S. (2003). Grensehandel og tax-free import av alkohol til Norge (English summary). *Nordisk Alkohol- & Narkotikatidsskrift*, **20**, 20-33.
- Nordlund, S. (2010). Alkoholbruk før og nå – har vår oppfatning endret seg? [Alcohol abuse in the past and the present – has our understanding changed?] In: Amundsen, E.J. (ed.) *Hva er misbruk og avhengighet? [What is abuse and dependence?]* Oslo: Norwegian Institute for Alcohol and Drug Research.
- Skog, O-J. (1985). The collectivity of drinking cultures: a theory of the distribution of alcohol consumption. *British Journal of Addiction*, **80**, 83-99.
- StataCorp (2009). *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP.
- Tolonen, H., Laatikainen, T., Helakorpi, S., Talala, K., Martelin, T., & Prattala, R. (2010). Marital status, educational level and household income explain part of the excess mortality of survey non-respondents. *European Journal of Epidemiology*, **25** (2), 69-76.
- Østhus, S. (2005). *Befolkningens holdninger til alkoholpolitikken [Public opinion on alcohol policy]*. Oslo: Norwegian Institute for Alcohol and Drug Research.

NOTES

- ¹ The data are made available in anonymous form by the Norwegian Social Science Data Services (NSD). Neither Statistics Norway nor NSD are in any way responsible for the analyses or interpretations done in the present study.
- ² The interviewers were instructed to record the time of the interview for administrative purposes, information which is included in the data files from Synovate Norway.
- ³ The 1995 sample was a household sample, and the response rate is not immediately comparable to other surveys.

- ⁴ In the 2008 HS, the last (highest) response category “4-7 times per week” was divided into two: “4-5 times per week” and “6-7 times per week”.
- ⁵ From the SUS data, two different measures of last year abstention are available: one defined as not having tasted beer, wine, liquor (or alcopops) during the past 12 months and the other defined as not having tasted any alcohol during the past 12 months. The estimated proportion of last year abstainers in the sample is virtually identical with these two measures.
- ⁶ In the 2004 and 2009 surveys, reported use of alcopops was also included in this measure. This has virtually no effect on the estimated proportion of abstainers or monthly drinkers in these years (a very small proportion of the samples have tasted alcopops).
- ⁷ Of course, 2-3 times per month does not correspond to 13 times per year, but 24-36 times per year. However, because the measure used could also be interpreted as “more than once per month” we have chosen 13 times per year as the cut-off point in the SUS measures. Note that the differences between the SUS and the HS surveys in the two last surveys would become larger, not smaller, if we instead had chosen 24 times per year as the cut-off point. In 1994/98 the proportion of monthly drinkers in Table 3 is higher in the SUS surveys than in the HS survey. Using 24 times per year would lower the SUS estimates and hence reduce the difference.
- ⁸ Because the NLF surveys are also (large) survey samples, the estimated proportion in paid employment from these data may also be misleading. However, due to the large samples (approximately 100 000 persons are surveyed each year) and high response rates, errors can be assumed to be small.