**Title:** The clinical consequences of variable selection in multiple regression models: a case study of the Norwegian Opioid Maintenance Treatment program

Words: 2749
Figures: 0
Tables: 2

**Running head title:** The consequence of variable selection

**Authors:** Marianne Riksheim Stavseth[1*], Thomas Clausen[1], and Jo Røislien[2]

[1] Norwegian Centre for Addiction Research, Institute of Clinical Medicine, University of Oslo, 0315 Oslo, Norway

[2] Faculty of Health Sciences, University of Stavanger, 4036 Stavanger, Norway

* Corresponding author: e-mail: m.r.stavseth@medisin.uio.no, phone: +47 92261179

## Abstract

**Background:** Selecting which variables to include in multiple regression models is a pervasive problem in medical research. The aim of this study is to compare the performance of different variable selection methods, and the potential clinical consequences of choice of method. The effect of missing data is also explored.

**Methods and material:** We used questionnaire data (n=18538, 69.9% men) from the Norwegian Opioid Maintenance Program. The dependent variable was engagement in criminal behavior while in treatment, and 29 potential covariates on demographics, psychosocial factors and drug use were tested for inclusion in a multiple logistic regression model. Both complete case and multiply imputed data were considered. We compared the results from variable selection methods ranging from expert-based and purposeful variable selection, through stepwise methods, to more recently developed penalized regression using the Least Absolute Shrinkage and Selection Operator (LASSO).

**Results:** The various variable selection methods resulted in regression models including from 9 to 22 covariates. The stepwise selection procedures generated the largest models. The choice of variable selection method directly affected the estimated regression coefficients, both in effect size and statistical significance. For several variables the expert-based approach disagreed with all data-driven methods.

**Conclusions**: The choice of variable selection method will strongly affect the resulting regression model, along with accompanying effect sizes and confidence intervals, thus affecting clinical conclusions. The process should consequently be given sufficient consideration in model building. We recommend combining expert knowledge with a data-driven variable selection method such as LASSO to explore the models' robustness.



**Keywords:** Logistic regression; Variable selection; Missing data; Opioid maintenance treatment; Crime

## 1. Introduction

The medical and social sciences are experiencing an enormous growth in data availability, and statistical modelling is becoming increasingly important to uncover structures within these data (1). When building a statistical model, one should aim for a model which is rich enough to capture relevant associations within the data, while simple enough to understand, interpret and use (2).

Variable selection, also referred to as subset or feature selection, is an important part of statistical model building. In perfectly designed experiments and randomized controlled trials, variables that might affect the outcome are controlled for by design, and few – if any – covariates need to be adjusted for in the accompanying statistical analyses (3). In observational studies however, a wide range of potential covariates are often measured, and the question of which ones to include in statistical models is an important one, as uncontrolled covariates may lead to biased results due to confounding (4).

Variable selection is the process of determining which variables to include in a statistical regression model. Finding the right balance between simplicity and richness is a key aspect of variable selection (5). If too few, or the wrong, covariates are included in the model, the model will not capture the essential structure in the data. The model is too simple, generally termed *underfitted*. Similarly, if too many, or the wrong, covariates are included in the model, the model will mistake random variation in the specific data set at hand for variation inherent in the general problem. The model ends up being too detailed, fitting the individual observations in the specific data set too well, generally termed *overfitted*. A good statistical model should be neither underfitted nor overfitted to the data at hand.

The logistic regression model is commonly used when exploring binary outcome variable (6). In medical and social science, the aim of a multiple logistic regression model is often to identify important associations or predictors of the outcome, both in terms of clinical and statistical significance. This identification process usually involves some kind of subset selection procedure (5). Purposeful Selection (PS) has become a standard method for variable selection in logistic regression (7). Used correctly, PS works well (8), but the method opens up for p-hacking, popularly referred to as ‘fishing’. Alternative approaches exist, including among others stepwise selection procedures based on objective mathematical criteria. Stepwise procedures are common (9), despite poor performance regarding p-values, biased standard errors and absolute values of regression coefficients (10-13). In later years

more mathematically robust methods have been developed, including penalized regression methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) (14).

While missing data is common in social and medical research, most research on variable selection ignores this issue by examining complete cases only, so called complete case (CC) analysis (15). Numerous studies do however discourage the use of CC (16-18). A widely accepted method for handling missing data is Multiple Imputation (MI) (19-22), but the practical and potential clinical consequences of performing variable selection on complete cases only versus on imputed data is rarely explored (15).

The aim of the present study is to illustrate the performance of various methods for variable selection in logistic regression models, with focus on readily available methods applicable to practical medical researchers. We compare the following five approaches, representing increasing levels of statistical sophistication; clinical expert-based variable selection, purposeful selection, stepwise regression based on Akaike's Information Criterion (AIC) (23) and the more restrictive Bayesian Information Criterion (BIC) (24), and penalized regression using LASSO. We further compare the results from CC analysis to those using MI.

The methods are tested on real data from the Norwegian opioid maintenance treatment (OMT) program. In a 2017 study, factors associated with criminal engagement while in OMT was explored (25). In the paper, a pre-selection of 13 potentially associated covariates was made based on clinical expertise. In the present study we compare these pre-selected covariates to those selected by various data-driven methods. Notably, the model based on the pre-selected variables from experts is the model least in concordance with any of the other models.

## 2. Material and methods

### *2.1 Setting*

The Norwegian Opioid Maintenance Treatment (OMT) program was initiated in 1998 as an optional treatment for persons with heroin use disorders (26). The program is evaluated annually using an assessment questionnaire monitoring the status of all OMT patients currently in the program. The questionnaire includes 49 questions regarding the patients' current treatment, employment and housing situation; information about medication type, dose, and urine testing regime; psychological situation; and frequency of drug and alcohol use during the last year and the last four weeks (27). There are also questions regarding the patients' treatment organization, but these were not included in the study.

*2.2 Study sample and measures*

A total of 18 538 annual assessment questionnaires collected in the 6-year period from 2005 through 2010 were available for the study. Reduction in criminal engagement among patients is an important goal of the OMT program and has been studied extensively in the literature (25, 28-30). Engagement in criminal activity while in treatment was used as the binary outcome variable in the current analyses, defined as whether the patient had been arrested, put in custody, been charged, and/or convicted of a crime within the last 12 months prior to the completion of the questionnaire, either self-reported or known to the staff. Twenty-nine variables from the questionnaire were included as potential explanatory variables.

*2.3 Ethics*

The collection and use of the data was granted the authors by The Norwegian Regional Committee for Medical Research Ethics (reference number 2012/1134).

*2.4 Variable selection methods*

There are numerous methods for variable selection in statistical model. In the present study we have focused on readily available and much-used approaches and automated statistical methods applicable for practical medical researchers. The methods are briefly described below.

*Expert-based selection:* In the expert-based approach for variable selection a subset of covariates is selected based on clinical experience and findings from the existing literature. While generally resulting in a clinically meaningful model, the approach runs the risk of including covariates that each are individually interesting, but also highly correlated, undermining the potential statistical significance of individual covariates (31). The approach also tends to favor already known associations: variables are included in the model exactly because they have previously been shown to be significant.

*Purposeful selection:* Purposeful Selection tests each covariate univariately against the outcome, and then includes the subset of variables with a univariate p-value below 0.25 for further extensive testing of the model (7). While potentially giving valuable insight into the data under study, the procedure is time-consuming as covariates must be taken in and out of the model manually with great attention to detail and often ad hoc trial and error using p-values as a guide. Other concerns are multiple comparisons (32) and ‘significance-chasing’ related to publication bias (33).

*Stepwise selection:* Forward inclusion starts with no variables in the regression model, adding one variable at the time according to a chosen model fit criterion whose inclusion gives the highest statistically significant improvement. Conversely, backward elimination starts with all covariates in the regression model, removing one variable at the time according to the chosen model fit criterion. We applied a combination of forward selection and backward elimination, often referred to as bidirectional elimination, using both AIC (23) and BIC (24) as optimization criteria. If we compare the two information criteria with respect to *consistency* and *efficiency*, which are classical themes in variable selection (34), generally AIC is not strongly consistent, though it is efficient, while the opposite is true for the BIC (2).

*LASSO:* LASSO (14) is a penalized regression method simultaneously allowing for both coefficient estimation and variable selection, by forcing coefficients to be zero when falling below a mathematically optimal threshold value (35). The forcing effect is determined by a regularization parameter, commonly optimized by cross-validation (CV) (36). We performed variable selection using LASSO using a 10-fold CV to optimize the regularization parameter.

*2.5 Missing data*

Of the 18 538 questionnaires available for analysis only 12 282 (66.3%) were complete. Both the outcome and covariates had missing values. Failing to handle missing data appropriately is problematic (37). Using complete cases only is the most common way to handle missing data in medicine and social sciences (20), however it implies removing observations from the data set and is known to reduce power and increase bias (21, 38). CC analysis has been argued against repeatedly in the statistical literature (16, 18, 39). A well-accepted method for handling missing data is multiple imputation (MI) (19). MI involves performing $m > 1$ independent imputations for all missing values, resulting in *m* complete datasets. The complete datasets are then analyzed individually using standard statistical methods and the results pooled together to one summary estimate (40). Based on a previous study on the effect of missing data in the OMT questionnaire data, the data was pre-processed using Multivariate Imputation by Chained Equations (MICE) (41) with 50 imputations. The missing data had a non-monotone pattern and was assumed to be Missing At Random (MAR) (17, 21).

In the analysis performed on complete cases, the included covariates were noted. For the imputed data, the percentage of times each covariate was chosen by the various variable selection methods in each of the imputed data sets was reported. Covariates selected in

more than 50% of the imputed datasets were considered as part of the final model. The different regression models selected by the various variable selection methods were then fitted, with estimated regression coefficients and corresponding confidence intervals. For the imputed data results were pooled.

### *2.6 Software*

All analysis were performed using R 3.3.1 (42). MI was performed using the function mice and the prediction matrix automatically generated using the `quickpred` function in the R package `mice` (41). All regression models in the expert-based approach and in PS were estimated using `glm` in R. Stepwise regression methods were performed using the function `stepAIC` with k = 2 for AIC and k = log(n) for BIC in the package `MASS` (43). Penalized regression was performed using the function `glmnet` with $\alpha = 1$ for LASSO in the package `glmnet` (44). The regularization parameter $\lambda$ in the penalized regression methods was estimates using the function `cv.glmnet` in `glmnet`. Correlations were calculated using the function `rcorr` in the package `Hmisc` (45).

## 3. Results

The median (range) correlation between all 29 covariates was 0.07 (0.00-0.72). The highest correlation was found between 'frequency of drug use' and 'daily functioning due to drug use' (r = 0.72), and 'frequency of drug use' and use of benzodiazepines (r = 0.65) and cannabis (r = 0.65).

The covariates included in the final model when using each of the five different variable selection methods for both complete cases (CC) only and multiply imputed (MI) datasets are shown in Table 1. The methods resulted in models including between nine (LASSO using CC only) and 22 (AIC using MI data) covariates. Six covariates were selected by all methods both when using CC only and MI data; age, gender, regional treatment center, living arrangement, supervised intake, and stimulant use. All data-driven methods, that is, all methods except the expert-based approach, included income and severe depression as relevant covariates. Reversely, alcohol use to intoxication, benzodiazepine, cannabis and opioid use last 4 weeks was included in the expert-based model, but in neither of the data-driven methods.

There were several differences for the same variable selection method when applied to complete cases only and multiply imputed datasets. Generally, more variables were selected in the MI setting compared to the CC setting. For example, stepwise regression based on AIC selected 17 covariates in the CC setting and 22 in the MI setting.

Most estimated adjusted Odds Ratios (aORs) and corresponding 95% confidence intervals (CI) were only marginally affected by the choice of variable selection method. However, for some variables the choice was crucial. For example, the estimated effect of living in an institution ranges from highly statistically significant using expert-based selection, with an aOR of 1.60 (CI: 1.15-2.25) using complete cases only and 1.70 (CI: 1.36-2.12) using multiply imputed data, to non-significant using PS, with an aOR of 0.95 (CI: 0.65-1.4) using CC and 0.98 (CI: 0.76-1.26) using MI data. Similarly, for the estimated effect of income. Using stepwise selection and BIC the aOR of living on social security benefits was 2.54 (CI: 1.64-4.10) when using CC and 1.35 (CI: 0.87-2.11) when using MI data. The aORs and 95% CIs for all covariates in all the multiple logistic regression model selected by the different variable selection methods is summarized in the Appendix.

## 4. Discussion and conclusions

The choice of variable selection method may significantly affect the clinical conclusions drawn from a multiple logistic regression model. Using real data from the Norwegian OMT program we have explored the effect of choice of variable selection method on which covariates are included in the final regression model, along with the accompanying estimated effect sizes and corresponding confidence intervals.

A distinctive feature of variable selection problems is their enormous size; even with moderate numbers of potential covariates, a search through all possible subsets of regression models will quickly become a daunting – if not impossible – task. Some reduction of the number of potential models is thus needed (46).

It might be tempting to add a long list of explanatory variables to be on the safe side. However most statistical models do not handle the presence of superfluous variables well; such variables will introduce noise and result in a model that is overfitted to the data (2). Also, the estimated effect of a specific covariate can be influenced by other variables included in the model. For example, a model including information on both income and occupation, two correlated variables, yielded a different estimated effect size of income compared to a model where occupation was excluded (OR for income changed from 1.21 to 1.65). Including too many, or too highly correlated, covariates in a regression model might thus affect clinical conclusion drawn from it.

When the data in this study was utilized in a study exploring factors associated with criminal engagement while in OMT treatment (25) a pre-selection of 13 covariates was selected

based on clinical expertise. The same expert-based model was applied here and compared to the various data-driven methods. Comparing these results, one might question whether the model from the expert-based selection indeed reflects the underlying structure in the data. Had the previously published study used a data-driven approach the conclusions in the paper might have been different.

While expert-based selection might seem reasonable from a clinical perspective, the approach tends to favor already known associations. Not only is this potentially problematic in terms of finding the actual underlying structure in the data under study, but the resulting model will thus lean towards a model that represents what the expert expects to find in the data, rather than letting the data speak freely. In balancing sufficient complexity with necessary simplicity, different data-driven variable selection methods can offer valuable assistance.

The present case study also demonstrates how the choice of using complete cases (CC) only rather than the more robust multiply imputed (MI) data will affect the variable selection process, and by that the final multiple regression model and the clinical conclusions drawn from it. CC analysis has repeatedly been argued against in the statistical literature, especially when the percentage of missing observations is high (17, 18). Variable selection and estimation on complete cases versus imputed data will influence results and performing MI should therefore be considered in the pre-processing of the data.

This study provides insight into the problem of model selection in a real-world setting, illustrating the practical and clinical consequences of choosing one variable selection method over another. Unfortunately, no single variable selection method is objectively the best; this will depend on the research question, previous research in the field, number of observations and potential covariates. Covariates selected by several variable selection methods demonstrate robustness to model choice. On the other hand, covariates that are included in some models but not in others, or variables that are significant in some models and not in others, should be interpreted with care.

We recommend that the choice of variable selection method is given sufficient consideration as part of the statistical model building. Rather than relying solely on one approach, we suggest combining clinical expertise with a data driven method such as LASSO in order to explore the robustness of the model and accompanying effect estimates and statistical significances.

## References


1. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. Journal of clinical epidemiology. 2016;71:76-85.
2. Claeskens G, Hjort NL. Model selection and model averaging: Cambridge University Press Cambridge; 2008.
3. Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schünemann H, et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev. 2011(4).
4. Egger M, Smith GD, Schneider M. Systematic reviews of observational studies. In: Egger M, Smith GD, Altman DG, editors. Systematic reviews in health care: meta-analysis in context: BMJ Publishing Group; 2001. p. 211-27.
5. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer Science & Business Media; 2003.
6. Hosmer DW, Lemeshow S. Logistic regression. In: Laake P, Lydersen S, Veierød MB, editors. Medical statistics in clinical and epidemiological research. Oslo: Gyldendal akademisk; 2012. p. 90-126.
7. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression, 3.ed. Hoboken, New Jersey: John Wiley & Sons; 2013.
8. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine. 2008;3(1):17.
9. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. European Journal of Epidemiology. 2009;24(12):733–6.
10. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3 ed. Philadelphia, USA: Lippincott Williams & Wilkins; 2008.
11. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2 ed: Springer International Publishing; 2015.
12. Greenland S. Modeling and variable selection in epidemiologic analysis. American journal of public health. 1989;79(3):340-9.
13. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology. 1992;45(2):265-82.
14. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267-88.
15. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27(17):3227-46.
16. King G, Honaker J, Joseph A, Scheve K, editors. List-wise deletion is evil: what to do about missing data in political science. Annual Meeting of the American Political Science Association, Boston; 1998.
17. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods. 2002;7(2):147.
18. Myers TA. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. Communication Methods and Measures. 2011;5(4):297-310.
19. Allison PD. Multiple Imputation for Missing Data. Sociological Methods & Research. 2000;28(3):301-9.
20. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiology. 2012;23(5):729-32.
21. Graham JW. Missing data analysis: Making it work in the real world. Annual review of psychology. 2009;60:549-76.
22. Van Buuren S. Flexible imputation of missing data. Boca Raton, FL: CRC press; 2012.
23. Akaike H. A new look at the statistical model identification. Automatic Control, IEEE Transactions on. 1974;19(6):716-23.
24. Schwarz G. Estimating the dimension of a model. The annals of statistics. 1978;6(2):461-4.
25. Stavseth MR, Røislien J, Bukten A, Clausen T. Factors associated with ongoing criminal engagement while in opioid maintenance treatment. J Subst Abuse Treat. 2017;77:52-6.

26. Waal H. Merits and problems in high-threshold methadone maintenance treatment. Eur Addict Res. 2007;13(2):66-73.
27. Riksheim M, Gossop M, Clausen T. From methadone to buprenorphine: Changes during a 10 year period within a national opioid maintenance treatment programme. J Subst Abuse Treat. 2014;46(3):291-4.
28. Bukten A, Skurtveit S, Gossop M, Waal H, Stangeland P, Havnes I, et al. Engagement with opioid maintenance treatment and reductions in crime: a longitudinal national cohort study. Addiction. 2012;107(2):393-9.
29. Bennett T, Holloway K, Farrington D. The statistical association between drug misuse and crime: A meta-analysis. Aggression and Violent Behavior. 2008;13(2):107-18.
30. Marel C, Mills KL, Darke S, Ross J, Slade T, Burns L, et al. Static and dynamic predictors of criminal involvement among people with heroin dependence: Findings from a 3-year longitudinal study. Drug and alcohol dependence. 2013;133(2):600-6.
31. Allen MP. Understanding regression analysis. New York: Plenum Press; 1997.
32. Greenland S. Multiple comparisons and association selection in general epidemiology. International Journal of Epidemiology. 2008;37(3):430-4.
33. Rothstein HR, Sutton AJ, Borenstein M. Publication bias in meta-analysis: Prevention, assessment and adjustments: John Wiley & Sons; 2005.
34. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: Springer; 2013.
35. Morozova O, Levina O, Uusküla A, Heimer R. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Medical Research Methodology. 2015;15(71).
36. Browne MW. Cross-Validation Methods. Journal of Mathematical Psychology. 2000;44(1):108-32.
37. Horton NJ, Kleinman KP. Much Ado About Nothing. The American Statistician. 2007;61(1):79-90.
38. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. New England Journal of Medicine. 2012;367(14):1355-60.
39. Allison PD. Missing Data. Thousand Oaks, CA: Sage 2001.
40. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 1987.
41. Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. Journal of statistical software. 2011;45(3).
42. R. R: A Language and Environment for Statistical Computing. 3.3.1 ed. Vienna, Austria: R Foundation for Statistical Computing; 2016.
43. Ripley B. Support Functions and Datasets for Venables and Ripley's MASS. 7.3-47 ed2017.
44. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1-22.
45. Harrell FE. Package 'Hmisc'. R Foundation for Statistical Computing. 2018.
46. George EI. The Variable Selection Problem. Journal of the American Statistical Association. 2000;95(452):1304-8.

**Table 1.** Covariates selected by the different variable selection methods in the complete cases and the percentage of times a covariate is selected by the different variable selection methods in the 50 imputed datasets.

| | | Complete Case | | | | Multiple Imputation | | | |
|---|---|---|---|---|---|---|---|---|---|
| | EB* | PS | AIC | BIC | LASSO | PS | AIC | BIC | LASSO |
| Age | x | x | x | x | x | 100 % | 100 % | 100 % | 100 % |
| Gender | x | x | x | x | x | 100 % | 100 % | 100 % | 100 % |
| Regional treatment center | x | x | x | x | x | 100 % | 100 % | 100 % | 100 % |
| Occupation | x | x | x | | | 100 % | 100 % | 90 % | 56 % |
| Income | | x | x | x | x | 100 % | 100 % | 100 % | 100 % |
| Work training | | | | | | 0 % | 0 % | 0 % | 0 % |
| Daily activity | | | | | | 22 % | 84 % | 0 % | 0 % |
| Living arrangement | x | x | x | x | x | 100 % | 100 % | 74 % | 100 % |
| Diagnosed Hepatitis C | | | | | | 44 % | 92 % | 8 % | 2 % |
| Diagnosed HIV | | | | | | 0 % | 18 % | 0 % | 0 % |
| Type of OMT medication | x | x | x | | | 100 % | 100 % | 100 % | 76 % |
| Prescribing doctor | | x | x | | | 56 % | 74 % | 0 % | 0 % |
| Prescription of benzodiazepines | | | x | | | 64 % | 94 % | 0 % | 0 % |
| Prescription of morphine-substance | | | | | | 0 % | 0 % | 0 % | 0 % |
| Nr. of supervised intakes per week | x | x | x | x | x | 100 % | 100 % | 100 % | 90 % |
| Severe anxiety (last 4 weeks) | | | | | | 0 % | 0 % | 2 % | 24 % |
| Severe depression (last 4 weeks) | | x | x | x | x | 100 % | 100 % | 96 % | 98 % |
| Severe delusion/hallucination (4 weeks) | | | | | | 18 % | 64 % | 6 % | 0 % |
| Severe somatic disease (last 4 weeks) | | | | | | 0 % | 0 % | 0 % | 0 % |
| Alcohol use to intoxication (last 4 weeks) | x | | | | | 0 % | 14 % | 0 % | 0 % |
| Benzodiazepine use (last 4 weeks) | x | | | | | 0 % | 58 % | 0 % | 64 % |
| Cannabis use (last 4 weeks) | x | | | | | 0 % | 8 % | 0 % | 2 % |
| Opioid use (last 4 weeks) | x | | | | | 0 % | 52 % | 0 % | 0 % |
| Stimulant use (last 4 weeks) | x | x | x | x | x | 100 % | 100 % | 100 % | 100 % |
| Frequency of drug misuse (last 4 weeks) | | x | x | | | 100 % | 100 % | 100 % | 4 % |
| Daily func.in reg. to drug misuse (4 weeks) | | x | x | | | 100 % | 100 % | 100 % | 100 % |
| Drug misuse (during last year) | x | x | x | x | x | 100 % | 100 % | 100 % | 0 % |
| Drug overdose (during last year) | | x | x | x | | 100 % | 100 % | 100 % | 0 % |
| Attempted suicide (during last year) | | x | | | | 68 % | 92 % | 8 % | 0 % |

*Expert-based variable selection is the same for CC and MI

# Appendix

Estimated adjusted Odds Ratios (aORs) with corresponding 95% confidence intervals for each model selected by the different variable selection methods.

| | Complete Case | | | | | Multiple Imputation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EB | PS | AIC | BIC | LASSO | EB | PS | AIC | BIC | LASSO |
| (Intercept) | 0.17 (0.10-0.29) | 0.04 (0.02-0.08) | 0.05 (0.02-0.11) | 0.03 (0.02-0.07) | 0.04 (0.02-0.08) | 0.19 (0.13-0.27) | 0.06 (0.03-0.11) | 0.06 (0.03-0.1) | 0.06 (0.03-0.11) | 0.13 (0.07-0.22) |
| Age | 0.95 (0.95-0.96) | 0.97 (0.96-0.98) | 0.97 (0.96-0.98) | 0.97 (0.96-0.98) | 0.96 (0.95-0.97) | 0.95 (0.95-0.96) | 0.96 (0.96-0.97) | 0.96 (0.96-0.97) | 0.96 (0.96-0.97) | 0.96 (0.95-0.97) |
| Gender: male | 1.78 (1.51-2.12) | 1.76 (1.48-2.1) | 1.75 (1.48-2.09) | 1.78 (1.5-2.12) | 1.74 (1.47-2.07) | 2.01 (1.78-2.27) | 1.98 (1.75-2.24) | 1.99 (1.76-2.26) | 1.99 (1.76-2.26) | 2.00 (1.77-2.26) |
| Centre: Buskerud | 1.15 (0.82-1.59) | 0.78 (0.53-1.12) | 0.76 (0.53-1.1) | 1.00 (0.71-1.38) | 1.04 (0.74-1.43) | 1.31 (1.01-1.69) | 1.03 (0.77-1.38) | 1.07 (0.8-1.43) | 1.15 (0.87-1.5) | 1.11 (0.86-1.44) |
| Centre: Midt | 0.86 (0.65-1.14) | 0.56 (0.4-0.78) | 0.54 (0.38-0.75) | 0.73 (0.55-0.96) | 0.72 (0.54-0.95) | 0.86 (0.68-1.09) | 0.66 (0.51-0.87) | 0.68 (0.52-0.89) | 0.75 (0.59-0.96) | 0.65 (0.51-0.82) |
| Centre: Øst | 1.6 (1.29-1.98) | 1.14 (0.89-1.47) | 1.11 (0.87-1.44) | 1.32 (1.06-1.64) | 1.35 (1.09-1.68) | 1.63 3 (1.4-1.9) | 1.33 (1.11-1.61) | 1.40 (1.16-1.69) | 1.39 (1.17-1.65) | 1.29 (1.11-1.51) |
| Centre: Telemark | 1.38 (1.00-1.9) | 1.15 (0.81-1.62) | 1.16 (0.82-1.63) | 1.3 (0.94-1.79) | 1.29 (0.93-1.79) | 1.58 (1.3-1.93) | 1.43 (1.14-1.79) | 1.45 (1.15-1.81) | 1.5 (1.22-1.85) | 1.37 (1.12-1.68) |
| Centre: Vestagder | 2.00 (1.53-2.62) | 1.89 (1.43-2.52) | 1.86 (1.39-2.48) | 1.9 (1.45-2.49) | 1.87 (1.42-2.45) | 2.02 (1.66-2.45) | 2.01 (1.63-2.49) | 2.03 (1.64-2.51) | 2.06 (1.67-2.54) | 1.76 (1.45-2.14) |
| Centre: Vestfold | 1.22 (0.86-1.71) | 0.71 (0.48-1.04) | 0.7 (0.48-1.03) | 0.90 (0.63-1.27) | 0.95 (0.67-1.34) | 1.58 (1.25-2) | 1.09 (0.83-1.44) | 1.12 (0.85-1.47) | 1.23 (0.96-1.58) | 1.17 (0.93-1.49) |
| Occupation: full time job | 0.4 (0.26-0.58) | 0.6 (0.35-1.02) | 0.59 59 (0.34-1) | | | 0.35 (0.26-0.48) | 0.44 (0.29-0.68) | 0.44 (0.29-0.67) | 0.45 (0.29-0.69) | 0.43 (0.28-0.65) |
| Occupation: part time job/student | 0.59 (0.45-0.77) | 0.71 (0.53-0.94) | 0.71 (0.53-0.93) | | | 0.6 (0.49-0.73) | 0.72 (0.59-0.89) | 0.73 (0.59-0.9) | 0.73 (0.59-0.9) | 0.68 (0.55-0.83) |
| Income: work assessment allowance | | 1.19 (0.67-2.18) | 1.21 (0.68-2.2) | 1.65 (1.08-2.62) | 1.65 (1.08-2.63) | | 0.86 (0.55-1.32) | 0.85 (0.55-1.32) | 0.85 (0.55-1.32) | 0.92 (0.6-1.43) |
| Income: disability or retire. pension | | 1.13 (0.63-2.09) | 1.14 (0.63-2.11) | 1.62 (1.06-2.59) | 1.64 (1.07-2.62) | | 0.91 (0.58-1.41) | 0.9 (0.58-1.41) | 0.89 (0.57-1.38) | 0.95 (0.61-1.48) |
| Income: social security benefits | | 1.86 (1.02-3.46) | 1.88 (1.04-3.5) | 2.54 (1.64-4.09) | 2.54 (1.64-4.10) | | 1.37 (0.88-2.15) | 1.36 (0.87-2.12) | 1.35 (0.87-2.11) | 1.43 (0.91-2.23) |
| Income: other | | 2.78 (1.48-5.33) | 2.78 (1.47-5.33) | 3.87 (2.31-6.64) | 3.95 (2.36-6.76) | | 2.11 (1.31-3.39) | 2.10 (1.3-3.37) | 2.09 (1.3-3.37) | 2.23 (1.39-3.58) |
| Daily activity | | | | | | | | 0.89 (0.78-1.02) | | |
| Living arr.: institution | 1.60 (1.15-2.25) | 0.95 (0.65-1.4) | 0.97 (0.66-1.42) | 1.39 (0.99-1.95) | 1.39 (0.99-1.95) | 1.70 (1.36-2.12) | 0.98 (0.76-1.26) | 0.98 (0.76-1.27) | 0.98 (0.76-1.27) | 1.43 (1.14-1.78) |
| Living arr.: stable | 0.65 (0.49-0.86) | 0.76 (0.57-1.03) | 0.78 (0.58-1.05) | 0.76 (0.57-1.02) | 0.74 (0.56-0.98) | 0.61 (0.5-0.74) | 0.75 (0.61-0.91) | 0.75 (0.61-0.91) | 0.76 (0.62-0.92) | 0.69 (0.57-0.84) |
| Living arr.: other | 1.04 (0.74-1.47) | 1.18 (0.83-1.68) | 1.22 (0.85-1.74) | 1.16 (0.82-1.64) | 1.15 (0.82-1.63) | 0.94 (0.74-1.19) | 1.1 (0.86-1.4) | 1.11 (0.87-1.42) | 1.11 (0.87-1.41) | 1.03 (0.81-1.3) |
| Diagnosed Hepatitis C | | | | | | | | 1.15 (0.99-1.32) | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| OMT medication: buprenorphine | 1.22 (1.06-1.41) | 1.19 (1.03-1.38) | 1.19 (1.03-1.38) | | | 1.30 (1.17-1.44) | 1.24 (1.12-1.38) | 1.26 (1.13-1.4) | 1.26 (1.13-1.4) | 1.30 (1.17-1.44) |
| Prescribing doc.: GP | | 1.29 (1.06-1.56) | 1.32 (1.09-1.61) | | | | 1.14 (0.99-1.31) | 1.15 (1.01-1.32) | | |
| Prescribing doc.: other | | 1.42 (0.85-2.3) | 1.34 (0.81-2.19) | | | | 1.12 (0.82-1.54) | 1.15 (0.84-1.58) | | |
| Prescribed Benzo.: yes | | | 0.86 (0.71-1.05) | | | | 0.88 (0.76-1.01) | 0.85 (0.74-0.98) | | |
| Nr. of supervised intakes pr. week | | 1.14 (1.1-1.19) | 1.15 (1.1-1.19) | 1.15 (1.11-1.19) | 1.15 (1.11-1.19) | | 1.12 (1.09-1.15) | 1.12 (1.09-1.15) | 1.12 (1.09-1.15) | 1.17 (1.14-1.2) |
| Depression: yes | | 1.37 (1.16-1.63) | 1.37 (1.15-1.62) | 1.36 (1.15-1.6) | 1.39 (1.18-1.64) | | 1.32 (1.16-1.5) | 1.34 (1.18-1.52) | 1.28 (1.13-1.45) | 1.32 (1.17-1.49) |
| Severe delusion/ hallucination: yes | | | | | | | | 0.88 (0.72-1.07) | | |
| Alcohol use to intoxication: yes | 1.15 (0.94-1.39) | | | | | 1.04 (0.91-1.19) | | | | |
| Benzodiazepine use: yes | 1.03 (0.86-1.23) | | | | | 1.01 (0.88-1.16) | | 1.11 (0.95-1.3) | | 1.08 (0.95-1.24) |
| Cannabis use: yes | 0.83 (0.71-0.98) | | | | | 0.89 (0.78-1.01) | | | | |
| Opioid use: yes | 1.02 (0.84-1.23) | | | | | 1.02 (0.89-1.18) | | 0.92 (0.79-1.06) | | |
| Stimulant use: yes | 1.67 (1.4-2) | 1.57 (1.31-1.88) | 1.56 (1.3-1.87) | 1.56 (1.32-1.84) | 1.58 (1.34-1.86) | 1.76 (1.54-2.01) | 1.68 (1.46-1.92) | 1.69 (1.47-1.94) | 1.67 (1.46-1.92) | 1.65 (1.44-1.88) |
| Daily functioning due to drug use: medium | 1.12 (0.92-1.37) | 1.21 (0.97-1.5) | 1.2 (0.97-1.5) | | | 1.13 (0.97-1.31) | 1.22 (1.05-1.43) | 1.21 (1.03-1.42) | 1.23 (1.05-1.43) | 1.3 (1.12-1.49) |
| Daily functioning due to drug use: low | 1.41 (1.1-1.81) | 1.65 (1.25-2.17) | 1.62 (1.23-2.15) | | | 1.56 (1.3-1.88) | 1.88 (1.53-2.3) | 1.86 (1.51-2.29) | 1.88 (1.54-2.31) | 1.95 (1.64-2.31) |
| Frequency of drug use: some use | | 0.83 (0.66-1.05) | 0.84 (0.66-1.06) | | | | 0.93 (0.79-1.11) | 0.9 (0.75-1.07) | 0.93 (0.79-1.11) | |
| Frequency of drug use: used regularly | | 0.58 (0.43-0.77) | 0.58 (0.44-0.78) | | | | 0.59 (0.48-0.72) | 0.56 (0.45-0.7) | 0.59 (0.48-0.72) | |
| Drug use: yes | 2.72 (2.17-3.42) | 2.36 (1.87-3.00) | 2.38 (1.88-3.03) | 2.29 (1.83-2.88) | 2.34 (1.87-2.93) | 2.1 (1.78-2.48) | 1.84 (1.55-2.19) | 1.85 (1.55-2.2) | 1.83 (1.54-2.17) | |
| Drug overdose: yes | | 2.53 (1.91-3.35) | 2.42 (1.84-3.18) | 2.59 (1.97-3.38) | | | 2.53 (2.07-3.09) | 2.52 (2.07-3.09) | 2.37 (1.96-2.87) | |
| Attempted suicide: yes | | 0.86 (0.62-1.19) | | | | | 0.79 (0.64-0.98) | 0.83 (0.67-1.04) | | |