Study Protocol

The association between tobacco use and COVID-19 infection and adverse outcomes in three Nordic countries: a pooled analysis

March 25, 2022

Revision history: this protocol has not been revised

Ahmed Nabil Shaaban[1], Sebastián Peña[3], Ida Henriette Caspersen[4], Filip Andersson[1,2] Sakari Karvonen[3], Per Magnus[4], Maria Rosaria Galanti[1,2]

1. Department of Global Public Health, Karolinska Institute, K9 Global folkhälsa, K9 GPH, 171 77 Stockholm, Sweden.

2. Centre for Epidemiology and Community Medicine, Stockholm Region, (CES), Solnavägen 1E (Torsplan), 113 65 Stockholm, Sweden

3. Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Mannerheimintie 166, PO BOX 30, 00271, Helsinki, Finland.

4. Centre for Fertility and Health, Norwegian Institute of Public Health, Postbox 222 Skøyen, N-0213 Oslo, Norway.

Funding:

The Tobrisk-CoV study is funded by NordForsk (project number 105544)

Background

The coronavirus disease (COVID-19) pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has caused more than 470 million confirmed cases of COVID-19, and more than 6 million deaths around the world by March 25, 2022 (WHO, 2022). Since the beginning of the COVID-19 pandemic, several factors have been attributed to increasing the risk of infection and adverse outcomes of the COVID-19 disease. Among these factors, the relationship between tobacco use and COVID-19 infection and adverse disease outcomes remained controversial as studies kept reporting mixed findings. Early studies reported what seemed to be a protective effect of tobacco use on COVID-19 infection (Haddad et al.; Jiménez-Ruiz et al., 2020), or hospitalizations due to COVID-19 (Farsalinos et al., 2020; Neira et al., 2021). A more recent ongoing living rapid review, this time including a larger selection of studies with different study designs, found that smokers are at reduced risk of SARS-COVID-19 infection compared to non-smokers (Relative risk 0.67, 95% Credible interval 0.60-0.75) (Simons et al., 2021). These findings opened the way for speculations and hypotheses on the potential mechanisms behind this protective role. However, results from most of these studies may be affected by selection bias as they reported findings from clinical samples or bias due to confounding as the structure of these published data permitted only univariate analysis.

Results from studies that suffer from selection bias or bias due to confounding should be handled with caution as they may undermine years of public health education against tobacco use, a major cause of morbidity and mortality worldwide. Moreover, the role of tobacco use in disease progression such as disease requiring hospitalization, ICU, and death remains unclear as most of the previous studies focused more on the association between tobacco use and the risk of infection, but not the adverse outcomes. These facts call for studies that ensure addressing any knowledge gap on the relation between tobacco and COVID-19 by taking into consideration 1) decreasing the risk for confounding and selection bias; 2) increasing precision

through a higher sample size, 3) further investigating the association between tobacco use and adverse disease outcomes. In most Nordic countries, the profile of tobacco use in the underlying populations allows the analysis of several types of tobacco use e.g. cigarette smoking and smokeless tobacco (snus) use, enabling further insights into the potential role of nicotine in the association between tobacco use and COVID-19. The use of smokeless tobacco is highly prevalent (even exceeding the prevalence of smoking among men in Sweden and Norway), which will allow us to disentangle a potential role of nicotine in the association between tobacco use and COVID-19.

The aim of this study is to examine the associations between tobacco use, COVID-19 infection, and adverse disease outcomes by using pooled population-based data from three Nordic countries, adjusting for potential confounders. The population-based nature of the samples minimizes selection bias Using a pooled analysis will accrue a large sample size and increase the potential for well–powered sub-groups analyses.

Methods

Setting and study design

This is an observational study of pooled population-based samples in three Nordic countries. Country-specific data has already been analysed in previous studies in Sweden (Galanti, 2021), Finland (Peña et al., 2021), and Norway (Magnus et al., 2016). We will use the national personal number assigned to every resident in these three countries at birth or at immigration to obtain information on diagnoses of COVID-19 and adverse outcomes among individuals in these cohorts through record-linkage with regional and national databases.

Data Sources

We will pool data from three Nordic countries, Sweden, Finland, and Norway. Briefly, Swedish data comes from a historical cohort of 424,386 clients of public dental clinics aged 23 and older in the Stockholm region with inception between October 2015 and January 2020, with follow-up from February 2020 to December 2020. In Sweden, the public dental clinics (Folktandvården, FTV) provide routine preventive visits (oral check-ups) to all residents who choose to receive care in these clinics. At each health check-up smoking and snus use are ascertained as past use, current use, and amount of current use. The national personal numbers assigned to every resident in Sweden at birth or at immigration will be used to obtain information on diagnoses of COVID-19 and of other diseases through record-linkage with regional health care registers. Demographic information will be extracted through record-linkage with the register of the total population of the Stockholm region held by Statistics Sweden.

The Finnish data will come from three pooled cross-sectional national health surveys in Finland (FinSote 2018-2020) of 44,199 participants aged 20 and older. The study samples included permanent residents in Finland from the FinSote surveys 2018, 2019, and 2020. The unique personal identifier assigned to all Finnish residents will be linked to the Communicable Diseases Registry to obtain information on diagnoses of COVID-19, to the Care Register for Health Care (HILMO) to obtain information on hospital admissions due to COVID-19, and to Statistics Finland Mortality Data to obtain information on deaths. Data on some sociodemographic characteristics will be also obtained from the Digital and Population Data Services Agency.

The Norwegian data will be based on the Norwegian Mother, Father and Child Cohort Study (MoBa) (Magnus et al., 2016), and the Norwegian Influenza Pregnancy Cohort (NorFlu) (Laake, 2018), with linkages to the Norwegian Surveillance System for Communicable Diseases (MSIS), the Norwegian Immunisation Registry (SYSVAK), and the Norwegian Population Registry. MoBa is a nation-wide population-based cohort consisting of 280 000 participants, where

parents were recruited during pregnancy from 1999 to 2008, while NorFlu is a pregnancy cohort consisting of 9 000 participants recruited in Oslo and Bergen during the swine flu pandemic in 2009-2010. Since March 2020, cohort participants have been regularly invited to answer electronic questionnaires. In June 2020 and January 2021, participants were asked to report current smoking. Questions about snus and other tobacco use were asked in January 2021 only. COVID-19 diagnoses are obtained from registry data (MSIS) based on PCR confirmed SARS-CoV-2 infection. Demographic information is extracted from the registries via linkage to the existing cohort databases. For the purpose of this study, all subjects who died before the onset of the pandemic (February 2020) in the three countries will be excluded from the analysis.

We developed a harmonization protocol to create a comparable dataset. The harmonization process started with identifying all possible variables to be used in the pooled analyses. We then identified common variables across countries, and only variables available in at least 2 countries will be included in the pooled analysis. Each country provided the format on which these variables were stored in their respective datasets. Finally, we elaborated a document describing the format for the harmonized variables. The principle of harmonization was to create variables with a minimum common format. The harmonized variables list is as follows:

Exposure Variables

Tobacco use (cigarette smoking and/or snus use) will be considered as exposure as follows:

1. *Current Cigarette Smoking*

Participants will be grouped as current cigarette smokers if they reported daily or occasional smoking. A categorical variable for current smoking status will be created (non-current smoker = 1, current daily or occasional smokers=2)

2. *Current amount cigarette smoked/day*

Current amount of cigarette smoked/day will be included as a continuous variable. Also a categorical variable of the originally reported average daily consumption will be created as follows: no smoking, 10 cigarettes per day (CPD) or less; 11-20 CPD; more than 20 CPD.

3. *Current Snus use*

Participants will be grouped as current snus users if they reported daily or occasional snus use. A categorical variable for current snus use will be created (non-current snus user= 1, current daily or occasional snus user= 2)

Outcome Variables

The study will include five outcomes: any diagnosis of COVID-19, hospital admission for COVID-19, length of hospital stay for COVID-19; intensive care use due to COVID-19, and death attributable to COVID-19. The follow-up period for all outcomes was from February 2020 until December 2020.

1. *Any diagnosis of COVID-19:*

Sweden: at least a positive polymerase chain reaction test (PCR) reported by the laboratories to Sweden's national electronic surveillance system for communicable diseases, SmiNet

Finland: cases with a positive SARS-CoV-2 RT-PCR, either informed by a laboratory or by a physician as a record of an ICD-10 code U07.1 (which requires a positive SARS-CoV-2 RT-PCR).

Norway: a positive test for SARS-CoV-2 based on PCR obtained from The Norwegian Surveillance System for Communicable Diseases (MSIS) or the presence of antibodies for SARS-CoV-2.

Out of these definitions a categorical variable will be created (No recorded COVID-19 diagnosis= 1, registered COVID-19 diagnosis= 2)

2. *Hospital admission for COVID-19*

Any hospital admission with a diagnosis of COVID-19 (ICD-10 codes U071 and U072). The diagnosis could be registered either as a main or as a concomitant diagnosis.

A. A categorical variable for hospital admission with any diagnosis of COVID-19 (either main or secondary diagnosis) will be created (No admissions= 1, any admission= 2).
B. A categorical variable for hospital admission with COVID-19 as the main diagnosis only (No admissions= 1, any admission= 2).

3. *Intensive unit care because of a diagnosis of COVID-19*

Admission to an intensive care unit (ICU) because of a diagnosis of COVID-19 (ICD-10 codes as above). A categorical variable for intensive unit care because of a diagnosis of COVID-19 (No/Yes) will be created (No ICU care= 1, any ICU care= 2).

4. *Death for COVID-19*

Death due to COVID-19 will be established using the Swedish, Finnish, and Norwegian Cause of Death Registries. All deaths occurring during the follow-up period with COVID-19 registered as the main cause will be included. The restriction to main cause will be done to maximize the specificity of the diagnosis (Ioannidis, 2021). A categorical variable for death due to COVID-19 will be created (Alive at the end of follow-up= 1, Death during follow-up= 2).

Covariates

We will include sex, age, cohabitation, education, employment, co-morbidity due to a tobacco-related disease and country of residence as covariates in the study. The covariates will be defined as follows:

1. *Sex*

Sex will be categorized as male (= 1), female (= 2).

2. *Age*

We will use age as a continuous variable (in years).

3. *Cohabitation*

We will define cohabitation as living alone (=1) or living with others (= 2).

4. *Education*

We will categorize education into three groups: Less than high school (= 1), high school (= 2), and university (= 3).

5. *Employment*

We will categorize current employment status as: no current employment, student or retired (= 1), currently part-time or full-time employed (= 2). This information is available in Sweden and Norway.

6. *Comorbidity due to any tobacco-related disease*

We will define any comorbidity due to a tobacco-related disease as: No tobacco-related comorbidity (= 1), any tobacco-related comorbidity (= 2). This information is available in Sweden and Norway.

7. *Country*

We will create a categorical variable with the country of residence as follows: Sweden (= 1), Norway (= 2), Finland (=3).

Date variables

1. *Year of latest assessment of cigarette smoking*

   Format: YYYY

2. *Year of latest assessment of snus use*

   Format: YYYY

3. *Date of COVID-19 Diagnosis*

   Format: DD/MM/YYYY

4. *Date of Hospital Admissions due to COVID-19*

   Format: DD/MM/YYYY

5. *Date of Intensive Care Admission due to COVID-19*

   Format: DD/MM/YYYY

6. Date of hospital discharge

   Format: DD/MM/YYYY

Statistical analysis

Risk ratios (RR) and their corresponding 95% CI for COVID-19 infection, hospitalization, ICU, or death due to COVID-19 will be estimated through generalized linear models (GLM). We will use the Poisson family with robust standard errors and the maximum likelihood optimization algorithm to estimate the relative risk of a confirmed COVID-19 case, hospital admission, ICU, or death due to COVID-19 (Zou, 2004). First, adjusted Poisson regression (Zou, 2004) will be applied to assess the influence of exposure (smoking or snus) on the study's outcomes. If we assume that y_{ij} is the observed binary outcome (COVID-19 infection, hospital admission, ICU, or death due to COVID-19) for subject i in country j, the specification and the equation of the Poisson model is as follows:

$$\log(p_{ij}) = \beta_0 + \beta_1 X_{1ij} + \sum_k^R \beta_k X_{ijk} \qquad (1)$$

where p_{ij} is the probability of the outcome y_{ij}, X_{1ij} is the exposure (smoking/snus), β_0 is the intercept, β_1 is the regression coefficient corresponding to the exposure, X_{ijk} represents the subject's values of R covariates, and β_k is a regression coefficient corresponding to each covariate. Exponentiating β_1 gives the corresponding relative risk, RR.

We will then apply a multilevel regression (Goldstein, 1995) which assumes that each country has its own COVID-19 infection probability, and this varies from one country to another. This will be the primary analytical model. In this multilevel model, the Poisson regression for COVID-19 infection will include an additional term u_j , which is the country-level random effects as a predictor variable:

$$\log(p_{ij}) = \beta_0 + \beta_1 X_{1ij} + \sum_k^R \beta_k X_{ijk} + u_j \qquad (2)$$

In this model, the probability depends on the value of the random effects u_j which is the totality of unmeasured country-level variables that predict COVID-19 infection and are uncorrelated with the individual variables in the model.

As a secondary analysis, we will use a negative binomial count model (Hilbe, 2011, 2014) to assess inpatient length of stay (i.e. the cumulative number of days of hospitalization) as a measure of disease severity among tobacco users compared to non-tobacco users.

We will carry out stratified analyses by sex, age, education, and time period.

We will carry out three sensitivity analyses. First, we will analyse the data adding tobacco-related comorbidity and employment as additional confounders with data from Sweden and Norway. Second, we will analyse only data prospectively collected to explore the influence of information bias. For this purpose, we will include data from Sweden and Finnish data from 2018 and 2019. This data will be analyzed with Cox proportional hazards models with the same specifications as the primary analysis. We will use time in study as the timescale. Participants will be censored due to end of follow-up or the occurrence of the event.

Third, for death of COVID-19, we will also include a competing risk analysis to take right-censoring due to a death non-attributable to COVID-19 into account. Competing risk means that a subject can experience one of a set of different events during the study period. Accordingly, the use of traditional methods of survival analysis, such as the Kaplan-Meier survival function, will result in incidence estimates that are biased upward, irrespective of the independence between the competing events (Austin et al., 2016; Berry et al., 2010; Kim, 2007; Southern et al., 2006). Instead, we will use two models to fit a regression in the presence of competing risk, the cause-specific hazard modeling and the sub-distribution hazard modeling (Austin et al., 2016). The cause-specific hazard function implies the instant rate of occurrence of a specific event in subjects who have not experienced yet any of the different types of events (Andersen et al., 2002; Austin et al., 2016). If we consider two types of events, death due to COVID-19 causes and death due to non-COVID-19 causes, then the cause-specific hazard of COVID-19 death implies the instant rate of COVID-19 death in subjects who have not yet experienced

either event (Austin et al., 2016). In the cause-specific hazards regression model, we will model the effect of covariates by using the traditional Cox proportional hazards model after censoring individuals with competing events at the time of their occurrence.

Unlike the cause-specific hazard function, the revised risk set in the sub-distribution hazard function (Fine & Gray, 1999), also known as the cumulative incidence function regression model, allows to include subjects who are event-free and subjects who already experienced a competing event . In the sub-distribution hazard function, we will model the effect of covariates by using a proportional hazards model through the cumulative incidence function that allows for the estimation of the incidence of the occurrence of an event while taking competing risks into consideration (Austin et al., 2016; Fine & Gray, 1999).

All analyses will be conducted with STATA®, version 17 (StataCorp LP, College Station, Texas, USA) and R version 4.1.2. We will incorporate the complex sampling design in Finland in all analyses.

References

Andersen, P. K., Abildstrom, S. Z., & Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical methods in medical research*, *11*(2), 203-215.

Austin, P. C., Lee, D. S., & Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, *133*(6), 601-609.

Berry, S. D., Ngo, L., Samelson, E. J., & Kiel, D. P. (2010). Competing risk of death: an important consideration in studies of older adults. *Journal of the American Geriatrics Society*, *58*(4), 783-787.

Farsalinos, K., Barbouni, A., & Niaura, R. (2020). Smoking, vaping and hospitalization for COVID-19. *Qeios*.

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, *94*(446), 496-509.

Galanti, M. R. (2021). Tobacco Use and the Risk of COVID-19. ClinicalTrials.gov: NCT04896918. 2021. [Available from: https://clinicaltrials.gov/ct2/show/NCT04896918].

Goldstein, H. (1995). Multilevel Statistical Models, Chapter 2. Edward Arnold. In: London, Wiley, New York.

Haddad, C., Malhab, S., Sacre, H., & Salameh, P. Smoking and covid-19: a scoping review. Tob Use Insights. 2021: 14: 1179173X21994612. In.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Ioannidis, J. (2021). Over-and under-estimation of COVID-19 deaths. *European journal of epidemiology*, *36*(6), 581-588.

Jiménez-Ruiz, C. A., López-Padilla, D., Alonso-Arroyo, A., Aleixandre-Benavent, R., Solano-Reina, S., & de Granda-Orive, J. I. (2020). COVID-19 and Smoking: A Systematic Review and Meta-Analysis of the Evidence. *Archivos de bronconeumologia*, *57*, 21-34.

Kim, H. T. (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical cancer research*, *13*(2), 559-565.

Laake, I. e. a. (2018). Risk of pregnancy complications and adverse birth outcomes after maternal A(H1N1)pdm09 influenza: a Norwegian population-based cohort study. BMC Infect Dis 18, 525, doi:10.1186/s12879-018-3435-8 (2018).

Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A. K., Handal, M., Haugen, M., Høiseth, G., & Knudsen, G. P. (2016). Cohort profile update: the Norwegian mother and child cohort study (MoBa). *International journal of epidemiology*, *45*(2), 382-388.

Neira, D. P., Watts, A., Seashore, J., Polychronopoulou, E., Kuo, Y.-F., & Sharma, G. (2021). Smoking and risk of COVID-19 hospitalization. *Respiratory medicine*, *182*, 106414.

Peña, S., Ilmarinen, K., Kestilä, L., & Karvonen, S. (2021). Tobacco Use and COVID-19 Incidence in the Finnish General Population (Tobrisk-CoV). ClinicalTrials.gov: NCT04915781. 2021. [Available from: https://clinicaltrials.gov/ct2/show/NCT04915781]. https://clinicaltrials.gov/ct2/show/NCT04915781

Simons, D., Shahab, L., Brown, J., & Perski, O. (2021). The association of smoking status with SARS-CoV-2 infection, hospitalisation and mortality from COVID-19: A living rapid evidence review with Bayesian meta-analyses (version 12). *Qeios*.

Southern, D. A., Faris, P. D., Brant, R., Galbraith, P. D., Norris, C. M., Knudtson, M. L., Ghali, W. A., & Investigators, A. (2006). Kaplan–Meier methods yielded misleading results in competing risk scenarios. *Journal of clinical epidemiology*, *59*(10), 1110-1114.

WHO. (2022). *Coronavirus (COVID-19) Dashboard.* https://covid19.who.int/]

Zou, G. (2004). A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, *159*(7), 702-706.