How did the second Covid-19 lockdown in Denmark affect tobacco purchases among frequent smokers?

A study analysing tobacco habits based on a time series of electronic receipts from various Danish supermarkets.

06 September 2021

Introduction

It is well-known that smoking is a risk factor for different diseases, cardiovascular disease amongst others [1]. 73 % of smokers in Denmark wish to quit, still, the smoking prevalence among adults in Denmark is 17% [2]. A pandemic is a radical societal intervention that affects the everyday life of people around the world, and the effect of the early phase of the COVID-19 pandemic on smoking cessation has been investigated in several countries. A Turkish study found that the COVID-19 outbreak was effective on smoking cessation [3], while the Wave 3 (2020) ITC Four Country Smoking and Vaping Survey including 6870 smokers conducted in Australia, Canada, England and the US showed that overall nearly 50% of the smokers thought about quitting, but only 14% actually made a quit attempt [4]. The thoughts of quitting have been related to the fear of COVID-19, justified by increasing evidence that smoking results in a graver COVID-19 disease progression [5]. So, the pandemic might offer a "leanable window" [6] or a "teachable moment" as they phrase it in [7], for smokers to change their behaviour. Therefore, understanding these behavioural changes due to the pandemic is important to obtain more focused and targeted interventions on smoking cessation.
Conversely, the pandemic also resulted in increased stress levels, as national lockdowns led to a series of restrictions being put into place, including closed schools and restaurants, physical distancing measures and remote workplaces. Many smokers report smoking as a way to cope with high levels of stress, and a recent study associated the initial COVID-19 lockdown in California with an increased cigarette consumption among current smokers [8]. A possible cause mentioned was an increased opportunity for smoking due to virtuel workplaces not being protected by indoor smokefree workplace laws, which underlines the importance of protective policies. Thus, understanding the public health effect of the COVID-19 pandemic and identifying exposed groups are important for developing and focusing targets for intervention. This is underlined in [9], where they conclude that there are socio-economic disparities in the association between smoking and COVID-19 infection based on a study of 53000 adults in the UK. In conclusion, there are conflicting results from various countries about how the early phase of the COVID-19 pandemic affected smoking cessation, and in all studies reviewed, the authors inform about limitations and possible improvements about the study design that one should be aware of when interpreting the results. In this paper, we suggest a novel approach

by using a time series of credit card transactions from 8800 subjects obtained in the period July 2018 to August 2021 from a large number of Danish supermarkets.

A novel approach: Tracking smoking behaviour over time using credit card transactions

To date, all existing literature concerning this topic, as described above, is based on questionnaires about smoking behaviour and thoughts about smoking and COVID-19. Common for these is that the questions are asked at one (or more) specific point in time, which does not allow for accurate tracking of changes over time [9]. During the pandemic, things changed so rapidly, that daily (even hourly) data is better suited to capture an accurate development, which the credit card transactions represent. This longitudinal data also gives the opportunity to investigate whether or not people actually return to their old habits again as time passes. A study using a similar type of data examined trends in the number of visitors, followers, and subscribers on smoking cessation digital platforms from January to April 2020, and compared these traffic data to a control period [10]. Here, they concluded that the initial increase in the number of visitors and subscribers was followed by a decline in traffic, and investigation of such trends is also possible using our credit card transaction data.
Secondly, we avoid recall bias and self-reported bias, which all existing studies suffer from to some degree. Only a few of the studies are based on real time data collection (before and after a COVID-19 lockdown) [8, 7], where the latter did not collect data in March 2020, and then changed from physical to phone interviews. In other studies, data about smoking habits before lockdown was collected retrospectively after lockdown, which induces a noticeable recall bias which might affect the results [9, 3]. Finally, we conduct our study in a Danish setting considering recent data from the second wave (lockdown in Denmark in January 2021-May 2021) which has not been seen yet in the literature. Limitations about our study design will be included in the discussion.

Supermarket transaction data structure

We have a large amount of supermarket transaction data, where each transaction is uniquely defined by person and time, and contains at least one item. In this framework, one can think of a transaction as a receipt. We let K denote the total number of transactions in the database. Letting M be the total number of items in the database, we define the set of items to be

$$\mathcal{I} = \{I_1, I_2, ..., I_M\}, \tag{1}$$

where item m is denoted I_m. Each item is associated with a positive item price and item quantity. This information is contained in the variables **item**, **itemprice** (in DKK) and **quantity** as seen in table 1 below. Here, the gray and white colors mark the different transactions, which are identified uniquely by a transaction id, **TID**. Thus, in below example in table 1, we have a database consisting of $K = 5$ transactions, and $M = 9$ items:

$$\mathcal{I} = \{I_1, ..., I_9\} = \{bread, dip, dressings, fresh\ eggs, apples, milk, wine, beef, yoghurt\}$$

Note that the total price of the transaction (**transactionprice**) is based on itemprice and quantity.

TID	person	time	item	itemprice	quantity	transactionprice
1	1	17-03-2019 08:03:00	bread	11.95	1	76.95
1	1	17-03-2019 08:03:00	dip	6.00	2	76.95
1	1	17-03-2019 08:03:00	dressings	53.00	1	76.95
2	1	19-03-2019 10:15:53	fresh eggs	27.95	1	78.40
2	1	19-03-2019 10:15:53	apples	15.00	0.700	78.40
2	1	19-03-2019 10:15:53	dip	10.00	2	78.40
2	1	19-03-2019 10:15:53	bread	19.95	1	78.40
3	2	02-02-2020 19:34:01	milk	9.95	1	9.95
4	2	14-02-2020 15:55:04	wine	109.00	3	479.80
4	2	14-02-2020 15:55:04	beef	49.95	2	479.80
4	2	14-02-2020 15:55:04	yoghurt	18.95	1	479.80
4	2	14-02-2020 15:55:04	bread	5.00	5	479.80
4	2	14-02-2020 15:55:04	milk	8.95	1	479.80
5	2	20-02-2020 20:24:10	apples	2.00	2	19.00
5	2	20-02-2020 20:24:10	bread	15.00	1	19.00
...	...	...	...	...	...	

Table 1: Transaction data example with 5 transactions and 9 different items.

- The structure of the transaction data is quite complex: each individual has multiple observations of transactions over time, and the transactions are quite irregular, meaning that the frequency of transactions differs over the weeks, months and between individuals.
- The number of items for each transaction will also vary between individuals and through time, as seen in above table. Furthermore, we will have new individuals entering the study and people dropping out, or even individuals dropping out and entering again, which creates missing time gaps.
- To describe this complex data structure we will adapt the theory of marked point processes [14] [15].

Supermarket transaction data: a marked point process

- Let T_k be the time for the k^{th} supermarket grocery transaction. For each transaction time, T_k, one or more items are purchased.
- For each transaction time T_k, we have information about the items in the transaction, which is described by $(X_1(T_k), ..., X_M(T_k))$. Specifically, $X_m(T_k) = (P_m(T_k), Q_m(T_k))$ denotes a vector that contain information about price and quantity for item m at time T_k.

- Consider the first transaction, $k = 1$, from the example in table 1. We have three different items $I_1 = bread$, $I_2 = dip$ and $I_3 = dressings$, with corresponding price and quantity. For bread we have:

$$X_1(T_1) = (P_1(T_1), Q_1(T_1)) = (11.95, 1) \in \mathcal{X}_1 = \mathbb{R}^+ \times \mathbb{R}^+$$

- Thus, for each transaction we have a positive real price and quantity for each item. We thus have a **mark space** for item m given by $\mathcal{X}_m = \mathbb{R}^+ \times \mathbb{R}^+$.
- With this example in mind, we can now define the marked point process, ϕ for the transaction times, T_k.

$$\phi = (T_k, (X_1(T_k), ..., X_M(T_k)))_{k \geq 1} \tag{2}$$

where each item has an associated mark space, such that $X_1(T_k) \in \mathcal{X}_1, ..., X_M(T_k) \in \mathcal{X}_M$. Note that the mark spaces do not depend on transaction, but only on item, and the mark space for the entire set of items $\mathcal{I}$ is given by $\mathcal{X}_1 \times ... \times \mathcal{X}_M$.

- Note that in most cases the quantities will be natural numbers, however, in some cases the quantity will be measured in kg or g, and the corresponding price will then be price per kg or price per g. As an example, see transaction two in table 1, where the costumer bought 0.7 kg apples that cost 15.00 DKK per kg.
- The defined marked point process from (2) counts the number of transactions made up to and including time t. When considering the process as a function of t, we have an integer-valued step function with jumps of size +1, which we assume to be right-continous, so that $N(t)$ is the number of events in the time interval $[0, t]$:

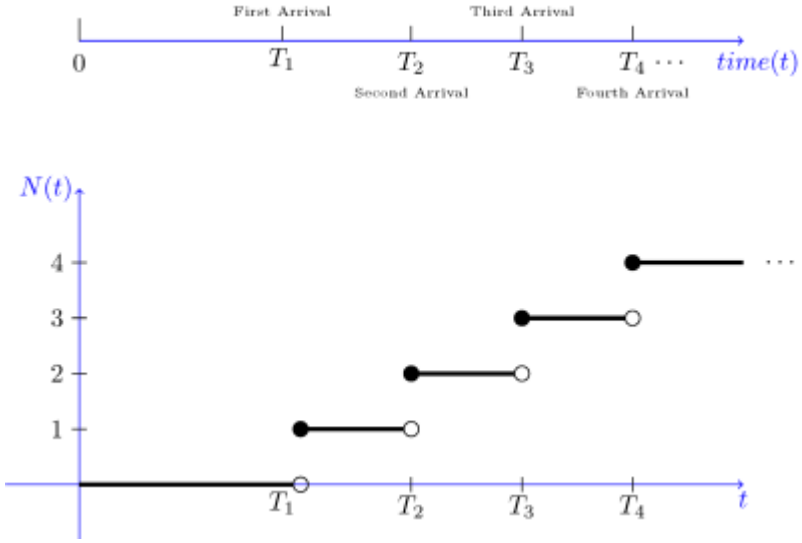
$$N(t) = \sum_{k \geq 1} I\{T_k \leq t\}, \tag{3}$$

where we assume $N(0) = 0$. See figure below for an illustration of the marked point process as a counting process [17].

Target parameter including examples

General target parameter

- We focus on a fixed time period $t \in [a, b]$ which is the same for every person and an item or itemset m^*. Note that m^* can be a specific item (example $m^* =$ tobacco) or an itemset (example $m^* =$ sugary drinks $=$ {soda with sugar, ice tea, alcohol free beer}).



- We now wish to define a target parameter in a general way, such that this can be used to investigate different questions. Following Appendix 4 (example A4.4 in Last and Brandt), we can write the target for m^* in the time period $[a, b]$ as a Lebesgue-Steiltjes integral. Here, we integrate with respect to the counting process defined above, as this process marks the arrivals of the transactions. We define:

$$\begin{aligned}\mu^{m^*}(a,b) &= E(\int_a^b f(X(t))dN(t)) \\ &= \sum_{k:\ a\leq T_k\leq b} E(f(X(T_k))),\end{aligned}$$

where $f(X(t))$ is a function defining the nature of the target. See below two examples for an understanding of this parameter.

Example 1

- We wish to investigate **the expected number of transactions containing m^* for** $t \in [a, b]$. Then we would define $f(X(t)) = I_{\{Q^{m^*}(t)>0,\ P^{m^*}(t)>0\}}$. So, $f(X(t))$ denotes the transactions where m^* was bought, ie. we have a positive quantity and price for m^*. From this, we could also calculate for example the expected number of daily or monthly transactions in the period.

Example 2 (maybe change notation here)

- The idea is to estimate the **expected number of item m^* for** $t \in [a, b]$.
- We use the same expression for $\mu^{m^*}(a, b)$, however, now defining $f(X(t)) = Q^{m^*}(t)$.

Including covariates and grocery shopping history

If we were to compare different groups or include grocery shopping history, we could condition on a filtration, $\mathcal{F}^-$, which denotes a set of possibly time varying covariates known just before time t as well as the grocery shopping history up until time t. Note that the shopping history is known at time T_k, but the time varying covariates can change at any time point, s. So, we have:

$$\mathcal{F}_{t^-} = \sigma\{Z(s) : s < t, T_k, X(T_k) : T_k < t\}$$

Using this, the upper example questions could be extended as follows:

Example 1 extension

- Question: "Are there any regions in which tobacco shopping changes during lockdown?" Translated to target: "For the five regions in Denmark, is the expected number of tobacco transactions in lockdown different from the preceding period"?
- Question: "Are there any regional differences in the change in tobacco shopping during lockdown?" Translated to target: "Does the change in the expected number of tobacco transactions between the lockdown period and the preceding period differ between regions?"

Example 2 extension

- Did the expected number of tobacco packages for males differ between lockdown and a control period?

Target parameter using intensities

First, we write up the target parameter for tobacco for the complete data, conditioning on information up to time t. Here, we use the tower property for conditional expectations. Let $N^{\text{tob}}(t) = N(t)Q^{\text{tob}}(t)$ be a counting process that counts the number of tobacco packages. We get the target parameter:

$$\begin{aligned}
\mu^{\text{tob}}(a,b) &= E(\int_a^b dN^{\text{tob}}(t)) \\
&= E\left[E(\int_a^b dN^{\text{tob}}(t)|\mathcal{F}_{t^-})\right] \\
&= E\left[\int_a^b E(dN^{\text{tob}}(t)|\mathcal{F}_{t^-})\right],
\end{aligned}$$

We now with to estimate this target based on the intensity of buying tobacco. From [12] we get the definition of an **intensity process** (here conditioning on $\mathcal{F}_{t^-}$):

$$\lambda^{\text{tob}}(t|\mathcal{F}_{t^-})dt = P(N^{\text{tob}} \text{ jumps in a time interval of length dt around time t } |\mathcal{F}_{t^-}) \quad (4)$$

where $\mathcal{F}_{t^-}$ (defined earlier) denotes the past up to the beginning of the small time interval dt (everything that has happened just before time t). In a small time interval, dt, N^{tob} either jumps once or does not jump at all. So the probability of a jump in that interval is close to the expected number of jumps in that interval [12]. From the definition of the intensity, we therefore have:

$$\lambda^{\text{tob}}(t|\mathcal{F}_{t^-})dt = E(dN^{\text{tob}}(t)|\mathcal{F}_{t^-})$$

Inserting this in the target parameter, we get:

$$\begin{aligned}\mu^{\text{tob}}(a,b) &= E\left[\int_a^b E(dN^{\text{tob}}(t)|\mathcal{F}_{it^-})\right] \\ &= E\left[\int_a^b \lambda^{\text{tob}}(t|\mathcal{F}_{t^-})dt\right]\end{aligned}$$

The observed data is not complete and consists of observations for subjects $i = 1, ..., n$, and our job is now to estimate the chosen target parameter based on this observed data.

Observed data: time scale and censoring

Choice of time scale

We choose to use calender time scale and not time since storebox start. By doing this, we will take seasonality into account and compare the same transaction times for different subjects. In this way, we can investigate our hypothesis about the impact of lockdown on tobacco shopping, as supposed to truncating the time scale by using time since storebox start, as people enter at different time points.

Censoring

In the observed data, we have different scenarios, which can lead to censored data. Examples are pictured in Figure 1. Here, we have shown transactions for five fictive subjects with and without tobacco in the period 1 Jan 2020 until 1 Jan 2021.

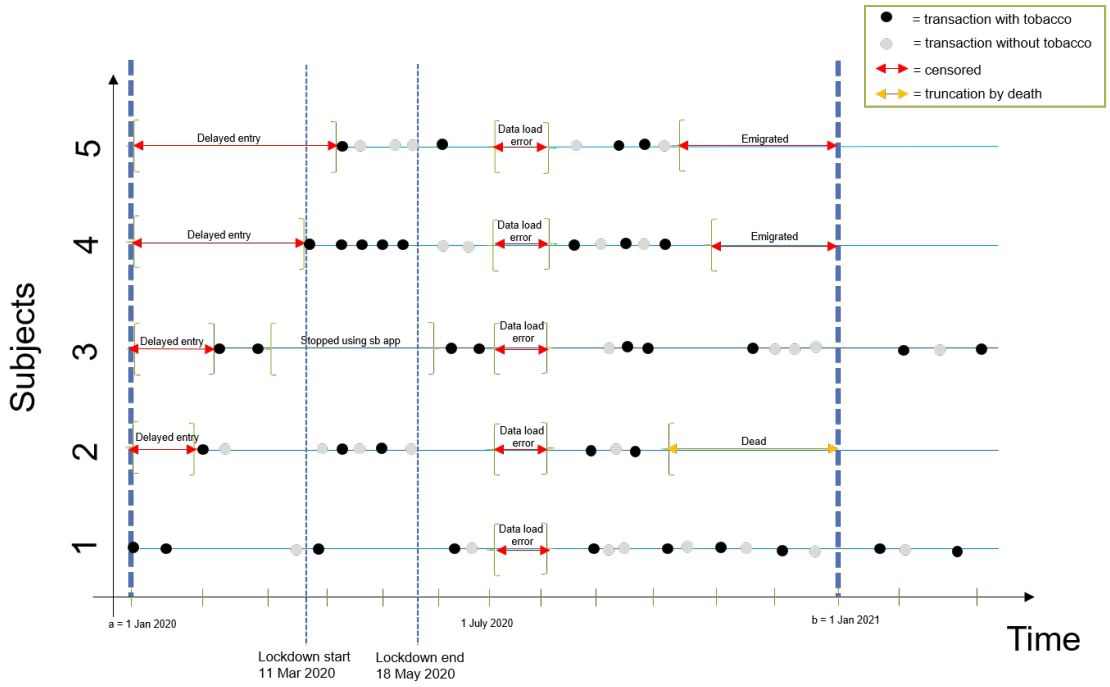


Figure 1: Missing mechanisms

Firstly, note that we have the case of truncation by death as marked by the orange arrow. In this case, the observations are not censored, as we know for sure that a transaction cannot happen from the grave! Secondly, in the periods with large gap times between transactions, we do not know whether the subject stopped using storebox, for example by changing to another supermarket (example subject 3), or if the subject realistically did not do frequent grocery shopping. For now we will not make any assumptions about these transaction gap times. In the periods marked by a red arrow, we know that the subject did not have a possibility of making a transaction using the storebox app. Therefore, the observed data is not complete as these periods are censored. Due to this censoring, $N(t)$ will not be fully observable, but only an incomplete version, $\tilde{N}_i(t)$ will be available for the i^{th} subject:

$$\tilde{N}_i(t) = N_i(t)C_i(t),$$

where the periods with censored data are delayed entry (caused by late entry into storebox or immigration), data load errors and emigration (assuming that the subject does not return to Denmark in the period $[a, b]$). So, we define the censoring process for the i^{th} subject as follows:

$$C_i(t) = \begin{cases} 0 & \text{if } \ a \leq t \leq \min(T_{i1}, b) & \text{(delayed entry)} \\ 0 & \text{if } \ e1^{\text{start}} \leq t \leq \min(e1^{\text{slut}}, b) & \text{(data load error)} \\ 0 & \text{if } \ e2_i^{\text{start}} \leq t \leq \min(e2_i^{\text{slut}}, b) & \text{(emigration)} \\ 1 & \text{otherwise} & \text{(transactions observed)} \end{cases},$$

where $e1_{\text{start}}, e1_{\text{slut}}$ denote the start and end dates for a data load error (same for all subjects) and $e2_i^{\text{start}}, e2_i^{\text{slut}}$ denote start end dates for emigration for subject i. The censoring process is shown in Figure 2 for subject 2 from Figure 1.

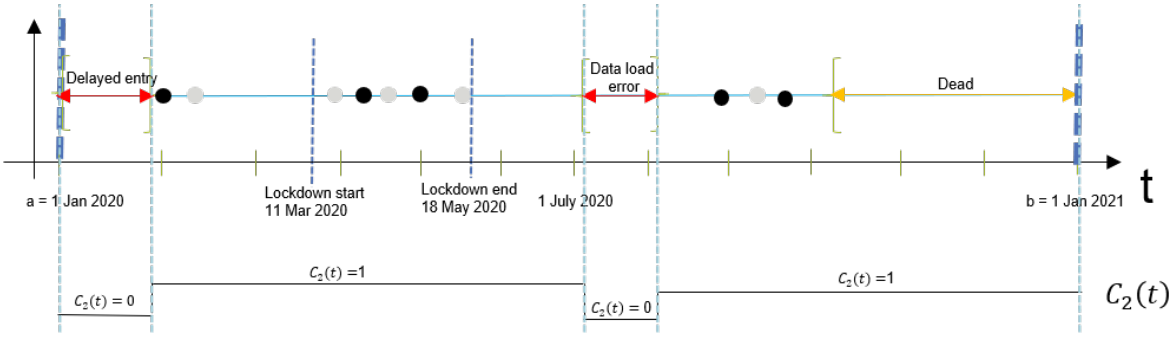


Figure 2: Example of the censoring process $C_2(t)$ for subject 2.

Thus, denoting the total number of transactions for subject i by K_i, we have the following observed data for subjects $i = 1, ..., n$ in the period $[a, b]$:

$$(C_i(t), \tilde{N}_i(t) : a \leq t \leq b, \max(T_{i1}, a) \leq t \leq \min(T_{iK_i}, b))_{i=1}^n$$

We will now write up an estimator for the target parameter based on this observed data.

Estimation of the target parameter

Assuming for now that expected number of tobacco transactions is independent of tobacco shopping history and covariates, we can remove the expectation, and thus wants to calculate:

$$\mu^{\text{tob}}(a, b) = \int_a^b \lambda^{\text{tob}}(t)dt$$

We now wish to investigate this target parameter in lockdown as compared to the preceding period (control period). Therefore, we define the following for the lockdown and control period:

$$L = [11.03.2020, 18.05.2020] \quad \text{and} \quad \overline{L} = [02.01.2020, 10.03.2020]$$

So, the null hypothesis that we wish to investigate is:

$$H_0 : \mu^{\text{tob}}(L) = \mu^{\text{tob}}(\overline{L})$$

that the expected number of tobacco packages in lockdown and the control period is the same. To estimate this based on the observed data, we start by plugging in an estimator for the intensity, as defined in [16] p. 77 (again assuming constant for $\mathcal{F}_{t^-}$):

$$\hat{\lambda}^{\text{tob}}(t) = \frac{\text{total number of tobacco transactions at time t}}{\text{total number of individuals at risk at time t}}$$

So, $\hat{\lambda}^{\text{tob}}(t)$ can be interpreted as the tobacco package rate per person-day, if we consider a time unit of days. So, we get:

$$\begin{aligned}\hat{\mu}^{\text{tob}}(a,b) &= \sum_{j\,:\,a\leq t_j\leq b} \hat{\lambda}^{\text{tob}}(t_j) \\ &= \sum_{j\,:\,a\leq t_j\leq b} \frac{d_j^{\text{tob}}}{n_j},\end{aligned}$$

where d_j^{tob} is the number of tobacco packages at time t_j and n_j is the total individuals at risk at time t_j (participants in storebox at time t_j).
In the lockdown and control period we get the following expected number of tobacco packages (assuming the subjects bought two packages at each transaction):

$$\hat{\mu}^{\text{tob}}(L) = \frac{2}{4} + \frac{2}{4} + \frac{6}{5} + \frac{2}{5} + \frac{4}{5} + \frac{2}{5} = 3.8 \;\; \text{pack/person}$$

$$\hat{\mu}^{\text{tob}}(\overline{L}) = \frac{2}{1} + \frac{2}{1} + \frac{2}{2} + \frac{2}{3} = 5.6 \;\; \text{pack/person}$$

- Consider to include the following points in the introduction:
- Data from hospital records (s. 9 uk inequelities)
- Lack of large sample sizes to be able to include confounders (spotlight).
- Differences at community level, not changes in specific participants' behaviour (p. 7 california).
- Research wise, an interesting perspective is that this turns an observational study into an interventional study.
- Stockpiling behaviour (letter).
- Not random sample (california), like in our case. Impacts generalizability. (p. 7)
- Impact on smoking cessation messages (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8083603/)

References

[1] FDA. (2020, April 5). How Smoking Affects Heart Health. Health Effects of Tobacco Use. https://www.fda.gov/tobacco-products/health-effects-tobacco-use/how-smoking-affects-heart-health.

[2] Sundhedsstyrelsen. (2018, March 6). Danskernes Sundhed - Den Nationale Sundhedsprofil 2017. https://www.sst.dk/da/udgivelser/2018/danskernes-sundhed-den-nationale-sundhedsprofil-2017

[3] Tetik, B. K., Tekinemre, I. G., Tas, S. (2021, June). *The Effect of the Covid-19 Pandemic on Smoking Cessation Success.*. Journal of Community Health. 46(3):471-475. doi: 10.1007/s10900-020-00880-2.

[4] Gravely, S., Craig, L. V., Cummings, K. M. et al. (2021, June 04). *Smokers' cognitive and behavioural reactions during the early phase of the COVID-19 pandemic: Findings from the 2020 ITC Four Country Smoking and Vaping Survey.* Plos One. 16(6):e0252427. doi: 10.1371/journal.pone.0252427.

[5] International Union Against Tuberculosis and Lung Disease (2021, August 17). *COVID-19 and TOBACCO: THE UNION'S BRIEF (Final Update: 17 August 2021).* https://theunion.org/our-work/covid-19/covid-19-and-smoking.

[6] Algatani, J. S., Aldhadir, A. M., Oyelade, T. et al. (2021, May 06). *Smoking cessation during COVID-19: the top to-do list.* NPJ Primary Care Respiratory Medicine. 31: 22. doi: 10.1038/s41533-021-00238-8.

[7] Jackson, E. J., Garnett, C., Shahab, L. et al. (2021, May). *Association of the COVID-19 lockdown with smoking, drinking and attempts to quit in England: an analysis of 2019-20 data.* Addiction. 116(5):1233-1244. doi: 10.1111/add.15295.

[8] Gonzalez, Epperson, A. E., Halpern-Felsher, B. et al. (2021, March 05). *Smokers Are More Likely to Smoke More after the COVID-19 California Lockdown Order.* Int J Environ Res Public Health. 18(5): 2582. doi: 10.3390/ijerph18052582.

[9] Jackson, E. J, Brown, J., Shahab, L., Steptoe, A., Fancourt, D. et al. (2020, August 21). *COVID-19, smoking and inequalities: a study of 53 002 adults in the UK.* Tobacco Control. doi: 10.1136/tobaccocontrol-2020-055933.

[10] El-Toukhy, S. (2021, March 22). *Insights From the SmokeFree.gov Initiative Regarding the Use of Smoking Cessation Digital Platforms During the COVID-19 Pandemic: Cross-sectional Trends Analysis Study.* Journal of medical internet research. doi: 10.2196/24593.

[11] Gill, D. R., Andersen, P. K. (1982). *Understanding Cox's Regression Model: A Martingale Approach*, The Annals of Statistics, 10 (4), p. 1100-1120.

[12] Gill, D. R. (1984). *Understanding Cox's Regression Model: A Martingale Approach*, Journal of the American Statistical Association, 79 (386), p. 441-447.

[13] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning* (Second ed.). Springer.

[14] Karr, A. F. (1986). *Point Processes and Their Statistical Inference* (First Ed.). Marcel Dekker.

[15] Last, G., Brandt, A. (1995). *Marked Point Processes on the Real Line* (First Ed.). Springer.

[16] Marubini, E. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies* (First Ed.). John Wiley and Sons.

[17] Pishro-Nik, H. (2014). *Introduction to Probability, Statistics, and Random Processes* (First ed.). Kappa Research.

[18] Tan, P. N., Steinbach, M., Karpatne, A., Kumar, V. (2019). *Introduction to Data Mining* (First ed.). Pearson.

[19] Walley, Rosalind et al. (2016). *Using Bayesian analysis in repeated preclinical in vivo studies for a more effective use of animals*, Pharmaceutical Statistics, No. 15, 2016, p. 277-285.