

מודל פשוט למעקב אחרי מספר המונשמים בישראל

אורי חפץ, עופר גליקסון, וגיא ישי¹

17 במאי, 2020

השאלה

הנחה מרכזית במאבק בנגיף הקורונה היא שיש להימנע מהמצב הנורא בו חולה קשה שזקוק להנשמה לא יוכל לקבל אותה. בישראל מדברים כיום על יותר מאלף מכונות הנשמה מצוותות וזמינות עבור חולי קורונה²? לפי נתוני משרד הבריאות³, שיא התפוסה עד היום היה באמצע אפריל, עם 140 חולי קורונה מונשמים. מספר המונשמים ירד מאז בהתמדה, אבל קיים חשש שעם פתיחת המשק במהלך חודש מאי, המספר יעלה שוב. **כיצד נוכל להתריע מראש על התקרבות למצב שבו נגמרות לנו המכונות (והצוותים)?**

המודל

כיום, כשיש לנו כבר מעל לחודשיים של נתוני מחלה יומיים בישראל, ניתן לאמוד מודל חיזוי סטטיסטי פשוט הקושר את מספר חולי הקורונה המונשמים ביום מסוים למספר חולי הקורונה המאומתים החדשים בימים שקדמו לו. המודל, המפורט בנספח המצורף, מתבסס על העובדה שכל חולה קורונה מונשם סווג לפני כן כחולה קורונה מאומת חדש. מתי זה "לפני כן"? מתישהו בין יום תחילת ההנשמה לבין מספר שבועות קודם. תחת מספר הנחות מפשטות לגבי יציבות לאורך זמן של אוכלוסיית הנדבקים והנדבקים ושל השפעות הנגיף, מתקבל קשר תאורטי לפיו מספר המונשמים ביום נתון ניתן לביטוי כפונקציה ליניארית של מספר הנדבקים המאומתים החדשים בימים הקודמים. יש להדגיש: אין מדובר במודל שחוזה את התפשטות המחלה. המודל קושר בין מדד אחד של התפשטות המחלה לאורך תקופה של מספר שבועות—מספר הנדבקים המאומתים החדשים—לבין מספר המונשמים בסוף התקופה.

כמו כל מודל מתמטי, איכות החיזוי של המודל תלויה באיכות ההנחות שבבסיסו. ההנחות שבבסיס המודל מפשטות מציאות מורכבת, ולכן אינן מדויקות. עד כמה תחזיות המודל שלנו מתקרבות למציאות?

חיזוי העבר

¹ האוניברסיטה העברית בירושלים

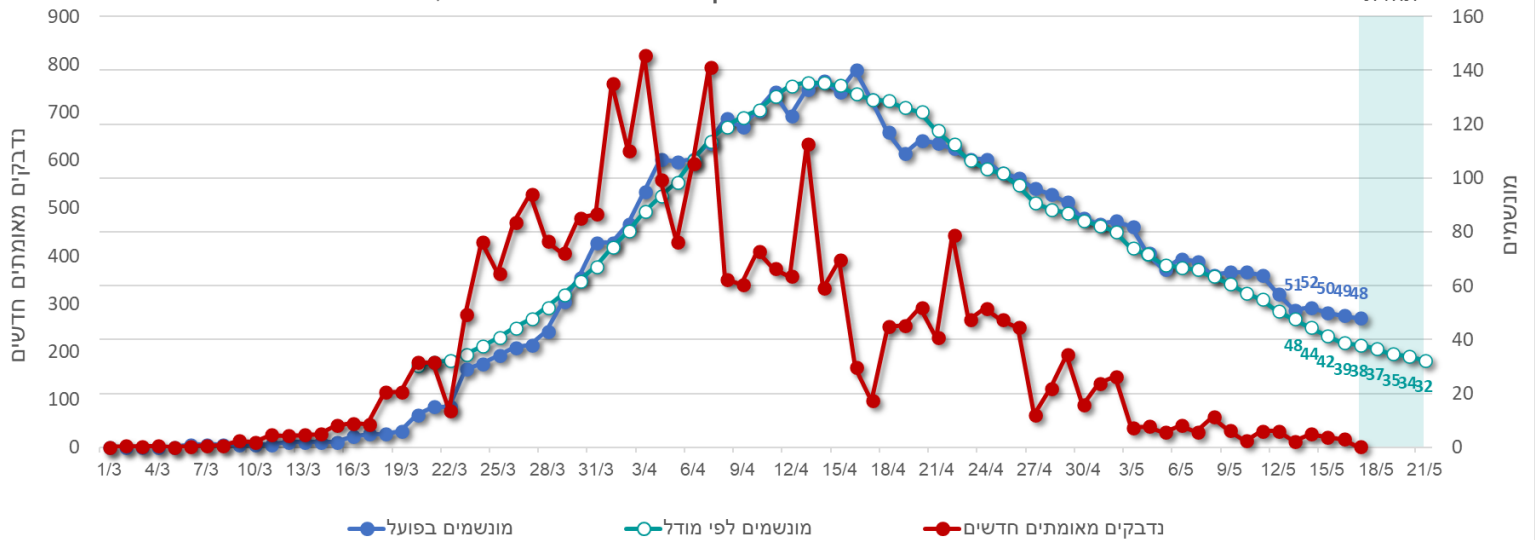
² נתונים בנושא ההתמודדות עם מגפת הקורונה, 26 למרץ 2020, מרכז המחקר והמידע, הכנסת

<https://main.knesset.gov.il/News/PressReleases/Documents/26032020.pdf>

³ מנהל מכלול אשפוז, מעודכן ליום 03/05/2020. נלקח מטלגרם משרד הבריאות <https://t.me/MOHreport>

כשבוחנים את התאמת המודל לנתוני העבר, רואים שהוא עושה עבודה מצוינת. מספר מונשמי הקורונה בישראל ביום נתון לחיזוי כמעט מושלם על סמך מספר חולי הקורונה המאומתים החדשים מהשבועות שקודמים לו. הגרף להלן ממחיש זאת.

מונשמים בפועל לעומת מודל ונדבקים מאומתים חדשים, ישראל



מבוסס על תוצאות הרגרסיה הליניארית לפרטים: אורי חפץ, עופר גליקסון וגיא ישי. מבוסס על נתוני מנהל מכלול האשפוז ותמונות מצב יומיות, מערכת הבריאות, 17/05/2020

העקומה האדומה מראה את כמות הנדבקים המאומתים החדשים בכל יום, החל מה-1 למרץ 2020 ועד היום, והכחולה מראה את מספר המונשמים (הכולל) בכל יום, החל מה-6 למרץ 2020 ועד היום, על פי נתוני מנהל מכלול האשפוז במשרד הבריאות⁴. העקומה התכולה מראה את תחזית המודל לגבי מספר המונשמים היומי. התחזית מבוססת על רגרסיה ליניארית של מספר המונשמים היומי על ארבעה ממוצעים של מספר הנדבקים החדשים ב-20 הימים האחרונים. הנוסחה המתקבלת עבור נתוני ה-17 במאי פשוטה מאוד (פרטים בנספח):

$$\text{מספר המונשמים ביום נתון} = 25 (\text{מספר קבוע})$$

- + 3% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים האחרונים (כולל היום)
- + 6% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים שלפניהם (לפני 5-9 ימים)
- + 7% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים שלפניהם (לפני 10-14 ימים)
- + 6% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים שלפניהם (לפני 15-19 ימים).

הנוסחה שקיבלנו מסתדרת עם הנתונים הרפואיים. בישראל מדברים כיום על כ-3-5 ימים מהידבקות במחלה ועד להופעת תסמינים (עבור אלו שמפתחים תסמינים), ועל כ-10 ימים

⁴ לפי הנתונים שבידינו, לא מפורט מספר המונשמים בין התאריכים 1 למרץ 2020 ועד ה-5 למרץ 2020. בתאריך ה-6 למרץ 2020 מספר המונשמים הוא 1.

מהיזבקות ועד להגעה למצב קשה (עבור אלו שמגיעים למצב קשה).⁵ כלומר, בהערכה גסה, מונשם שאובחן מאוחר—נניח, רק עם הגיעו לבית החולים במצב קשה—מדווח כנדבק חדש ביום תחילת היותו מונשם. מונשמים שאובחנו מוקדם יותר—נניח, יומיים אחרי שהחלו להופיע תסמינים—דווחו כנדבקים חדשים עד כ-5 ימים קודם. ומכיוון שמשך ההנשמה הוא כשבוע עד שבועיים, מרבית המונשמים ביום נתון דווחו כנדבקים חדשים מתישהו ב-20 הימים שמסתיימים ביום האמור.

מה לגבי המשקלות? נניח שבכל יום 100 נדבקים מאומתים חדשים. כעבור 20 יום, המודל חוזה 21 מונשמים (6+7+6+3 פחות 1 בשל עיגולים) מעבר למספר הקבוע (25). 21 מונשמים הוא גם המספר שנקבל אם הסיכוי הממוצע של נדבק מאומת להיות מונשם הוא 2.1% ומשך ההנשמה הממוצע הוא 10 ימים—מספרים שמסתדרים גם הם עם הערכות המומחים. ומה עם הקבוע 25? אנו מראים בנספח התיאורטי שהוא מרמז על טעויות במודל (למשל, יש מונשמים שאומתו כנדבקים לפני 20 הימים האחרונים) או בנתונים. העובדה שאינו גדול במיוחד ביחס למספר המונשמים הממוצע בחודשיים האחרונים מצביעה על כך שהטעות אינה גדולה.

חיזוי העתיד?

הנוסחה לעיל פשוטה, והיא מתבססת על נתונים פשוטים שמתפרסמים בכל יום באמצעי התקשורת בישראל. היא מתארת טוב מתאם סטטיסטי בין נתונים היסטוריים. אבל האם ניתן להשתמש בה על מנת לחזות את העתיד?

התשובה הזוהירה היא לא, אלא אם כן מה שהיה הוא שיהיה. כלומר, לא מומלץ להשתמש במודל כזה בשביל לחזות את העתיד אלא אם כן מניחים שהמתאם הסטטיסטי ההיסטורי שהגרף מראה ימשיך להחזיק גם בעתיד. תחת ההנחה הזו, הגרף מראה את מספר המונשמים החזוי בארבעת הימים הבאים. (מכיוון שאיננו יודעים את מספר הנדבקים המאומתים החדשים בארבעת הימים הבאים, החיזוי משמיט מהנוסחה לעיל עד ארבע תצפיות: של מחר, מחרתיים, וכו').

התחזית שהגרף מראה היא למטרות הדגמה בלבד. אפשר כמובן להאריך אותה כמה שרוצים—אבל עבור כל יום נוסף של חיזוי, נצטרך להחליף נתוני נדבקים ממשיים בהנחה או בתחזית לגבי נתונים עתידיים. כפי שחישבנו לעיל שעם 100 נדבקים מאומתים חדשים בכל יום, המודל חוזה 21+25 מונשמים החל מהיום ה-20, ניתן לחשב שעם 1,000 נדבקים מאומתים חדשים ביום, המודל חוזה 210+25 מונשמים החל מהיום ה-20.

אפשר לשאול שאלות מעניינות יותר, כמו: כמה מונשמים המודל חוזה ביום ה-20 אם נתחיל עם 100 נדבקים מאומתים חדשים ביום הראשון, ונגדיל את המספר ב-10% בכל יום? במקרה כזה סך הנדבקים המאומתים החדשים ביום ה-20 יגיע ל-612, והמודל ינבא שכמות המונשמים ביום זה תהיה 78. הטבלה להלן מחשבת את חיזוי המודל תחת 25 תרחישים כאלו. בפניה השמאלית התחתונה שלה, הטבלה מראה שגם אם ביום הראשון יש 400 נדבקים מאומתים

⁵ ראו, למשל, "המאבק בקורונה בראי צוות המומחים המייעץ למל"ל", אלי וקסמן, 5 למאי, 2020. <https://www.academy.ac.il/SystemFiles2015/Weizmann%20talk%20May%202020.pdf>

חדשים, וגם אם המספר יגדל בקצב מהיר של 20% ביום, ביום ה-20 יהיו בישראל 623 מונשמים, כלומר משמעותית פחות ממספר מכונות ההנשמה הזמינות. (לשם השוואה, אתמול, כלומר ב-16 למאי, היו בישראל פחות מ-10 נדבקים מאומתים חדשים).

טבלת ניתוח רגישות למספר המונשמים החזוי ביום 20						
שיעור הגידול היומי של נבדקים חדשים מאומתים						
20%	15%	10%	5%	0%		
100	69	51	41	36	50	מספר הנבדקים המאומתים החדשים ביום 1
175	113	78	58	46	100	
324	200	130	90	67	200	
474	288	182	122	88	300	
623	375	235	155	109	400	

אז בינתיים, נראה שמצבנו טוב. אבל צריך כמובן להיזהר: לא היו בישראל ימים עם 1,000 נדבקים מאומתים חדשים, ולא היו בישראל רצפים של 20 יום שמתחילים ב-400 מאומתים חדשים וממשיכים עם קצב גידול של 20%. אם הרכב האוכלוסייה הנדבקת או הנבדקת בתרחישים כאלה יהיה שונה מהותית מזה שהיה בישראל בחודשיים האחרונים, המודל ייכשל בניבוי. בשביל ניבוי אמין יותר, צריך מודלים עשירים יותר מהמודל הזה. המודלים שמשמשים את משרד הבריאות, המל"ל, ואלו שמתפרסמים על ידי צוותים של אפידמיולוגים, רופאים, מתמטיקאים ופיסיקאים, כולל בישראל, מתבססים על הרבה יותר מידע. הם לוקחים בחשבון נתונים נוספים, שהחשוב בהם הוא גיל האוכלוסיות השונות (הנדבקים, הנבדקים החדשים, והמונשמים). החיזויים של מודלים כאלה צפויים להיות טובים יותר, אבל הם קשים יותר להבנה בידי הציבור, והם מסתמכים על נתונים רבים יותר, כולל כאלה שאינם זמינים לציבור בקלות. אנחנו מקווים שיש ערך לצידם גם במודל הפשוט שלנו, ושהבנתו תעזור לתת לציבור תחושת שליטה, הבנה, וביטחון.

סיכום

ניתן להסביר את מספר המונשמים בכל יום נתון בחודשיים האחרונים כפונקציה פשוטה של מספר הנדבקים המאומתים בשלושת השבועות שלפניו. אם נמשיך את הבדיקות כבעבר, ואם הקשר הסטטיסטי יימשך, נוכל לדעת מבעוד מועד אם אנחנו מתקרבים לתפוסה מלאה של מכונות ההנשמה. כרגע, בכל אופן, אנחנו מאוד רחוקים (התפוסה כרגע היא של פחות מ-5% מהמכונות הזמינות). צריך להמשיך להיזהר, אבל כרגע אפשר לישון בשקט בלילה.

A Model of Covid-19 Ventilation-Patient Numbers - Appendix

Ofer Glicksohn

Ori Heffetz

Guy Ishai

May 20, 2020

1 The Model

Our goal is to explain and predict the number of ventilation patients in a given day t due to covid-19, using data on newly confirmed cases in the days prior to day t .

Denote a time interval with covid-19 outcomes on which we focus by \mathcal{T} . Denote the number of ventilation patients in day $t \in \mathcal{T}$ by V_t . Ventilation is an uncertain outcome for those who are newly confirmed with covid-19. To formalize this uncertainty, define the probability that a newly confirmed patient would need ventilation at day T after confirmation as p_T . Suppose that this probability becomes effectively zero when T is larger than a threshold \bar{T} , as a newly confirmed patient either needs ventilation or not within a limited time interval, and those who become ventilated patients eventually either recover or die. This leaves us with positive probabilities only up to this maximum-relevant-time-window threshold. Denote the number of newly confirmed patients in day $t \in \mathcal{T}$ by I_t and the vector of numbers of newly confirmed patients on t and the relevant previous days $\mathbf{I}_t = (I_t, \dots, I_{t-\bar{T}})$.

Based on ventilation probabilities, the number of newly confirmed patients in day $t - T$ who are ventilated in day t , i.e., the *contribution* of day $t - T$ to the number of ventilation patients in day t , is a random variable $V_{t,t-T}$. This random variable has the following expectation conditional on I_{t-T} :

$$E(V_{t,t-T} | I_{t-T}) = p_T I_{t-T}.$$

The total number of ventilation patients in day t is the sum of these random variables with $p_T > 0$,

$$V_t = \sum_{T=0}^{\bar{T}} V_{t,t-T}.$$

Therefore:

$$V_t = \sum_{T=0}^{\bar{T}} p_T I_{t-T} + \varepsilon_t, \quad (1)$$

where $\varepsilon_t = \sum_{T=0}^{\bar{T}} (V_{t,t-T} - p_T I_{t-T})$ is a random variable equal to the sum of deviations of the random variables $V_{t,t-T}$ from their expectations. Therefore

$$E(\varepsilon_t | \mathbf{I}_t) = 0$$

holds by definition in the period of interest. Equation 1 shows a linear relation between the number of ventilation patients in day t and the numbers of newly confirmed patients in prior days, given a mean-zero error term.

1.1 An ideal regression model

Set the time interval of focus as the sampling period, $\mathcal{T} = \mathcal{T}_{\text{sample}}$. The model in equation 1 can be estimated using a Generalized Least Squares (GLS) regression, and since by definition $E(\varepsilon_t | \mathbf{I}_t) = 0$, each \hat{p}_T is an unbiased and consistent estimator of the probability p_T relevant to the sampling period. Since the error terms may be serially correlated, a non-trivial error matrix needs to be chosen to obtain reliable estimates of the standard errors (see details in the empirical section).

The sample-relevant p_T probabilities take into account all the factors contributing to the uncertainty of ventilation use in the sample—variation in tests and diagnoses (e.g., who is tested, test quality), variation in confirmed-infected populations (e.g., outbreaks among different demographic groups), variation in the virus’s behavior, all sorts of measurement errors, and so forth—and provide an aggregate representation of this uncertainty. The estimated \hat{p}_T ’s are interpreted as the *empirical* probabilities relevant to the sampling period.

1.2 The benchmark regression model

In practice, due to data constraints, we have to limit the length of the relevant time window \bar{T} . This means that when estimating the model we omit some lagged I_{t-T} variables with $p_T > 0$. Suppose now that $p_T > 0$ is allowed for some $T > \bar{T}$ values. Since the I_{t-T} variables with $T \leq \bar{T}$ may be correlated with those with $T > \bar{T}$, estimates are potentially biased now (due to omitted variable bias).

To understand this bias, define $V_{t,\infty} = \sum_{T>\bar{T}} V_{t,t-T}$ and $I_{t,\infty} = \sum_{T>\bar{T}} I_{t,t-T}$. Define the overall probability that a patient who is newly confirmed in one of the days prior to \bar{T} is

ventilated in day t by p_∞ . Therefore

$$p_\infty = \frac{1}{I_{t,\infty}} \sum_{T>\bar{T}} p_T I_{t-T},$$

and it follows that $E(V_{t,\infty}|I_{t,\infty}) = p_\infty I_{t,\infty}$. Begin with the full model with all time lags:

$$V_t = \sum_{T=0}^{\bar{T}} p_T I_{t-T} + p_\infty I_{t,\infty} + \tilde{\varepsilon}_t, \quad (2)$$

where $\tilde{\varepsilon}_t = \varepsilon_t + (V_{t,\infty} - p_\infty I_{t,\infty})$ and $E(\tilde{\varepsilon}_t | (\mathbf{I}_t, I_{t,\infty})) = 0$. It follows that $\tilde{\varepsilon}_t$ is uncorrelated with the relevant numbers of confirmed-infected patients: $E(\tilde{\varepsilon}_t \mathbf{I}_t) = 0$. Now suppose that $I_{t,\infty}$ is potentially correlated with each I_{t-T} variable for $0 < T \leq \bar{T}$, such that the following linear projection can be estimated based on the sample:

$$I_{t,\infty} = \beta + \sum_{T=0}^{\bar{T}} \gamma_T I_{t-T} + \nu_t,$$

where $E(\nu_t \mathbf{I}_t) = \mathbf{0}$. The constant β is the best within-sample prediction of the part of $I_{t,\infty}$ that is not captured by the correlations. In the extreme (and unlikely) case of no correlations β is just the sample mean of $I_{t,\infty}$. Substituting in equation 2 we derive

$$V_t = p_\infty \beta + \sum_{T=0}^{\bar{T}} (p_T + p_\infty \gamma_T) I_{t-T} + (p_\infty \nu_t + \tilde{\varepsilon}_t).$$

Denote

$$\alpha = p_\infty \beta,$$

$$\pi_T = p_T + p_\infty \gamma_T,$$

and $e_t = p_\infty \nu_t + \tilde{\varepsilon}_t$. It follows that $E(e_t \mathbf{I}_t) = 0$. Therefore, estimating the regression model

$$V_t = \alpha + \sum_{T=0}^{\bar{T}} \pi_T I_{t-T} + e_t \quad (3)$$

provides consistent estimates for α and π_T .

The estimated $\hat{\pi}_T$'s can no longer be interpreted as proxies of the probabilities p_T in the sampling period. Instead, they now *overestimate* or *underestimate* these probabilities, depending on the relation between the number of confirmed infections I_{t-T} and the omitted number of confirmed infections I_∞ , and on the aggregate probability of the omitted confirmed

infections to become ventilation patients at day t , p_∞ . Assuming that the serial correlation of confirmed infections diminishes with the time gap between two days, this bias should be the most severe for the T values closest to \bar{T} . The bias should decrease when a larger \bar{T} is chosen.

The estimated $\hat{\alpha}$ is interpreted as a proxy for the *mean number of patients inside the sampling period* who need ventilation beyond \bar{T} days after confirmation, which is *not already explained* by the I_{t-T} variables based on their correlation with I_∞ . A large sample and a sufficiently large \bar{T} should yield $\hat{\alpha}$ close to zero, but estimating $\hat{\alpha}$ close to zero does not necessarily indicate that we account for a large enough \bar{T} . (It could mean, for example, that the autocorrelation of the confirmed infections time series is strong.)

Does the model make sense *inside* the sampling period? We now turn from identification issues to the more important question in our context: does the model of *fixed* probabilities p_T (plus a constant) provide an accurate representation of reality, or an inaccurate abstraction? This question is important both if the model is estimated with biases or without them.

When examining the value of the model inside the sampling period, we focus on R^2 , on the residuals and on $\hat{\alpha}$. An R^2 close to 1 and a non-systematic pattern of residuals suggest a large explanatory power of the model. An estimated parameter $\hat{\alpha}$ close to zero, relative to the typical numbers of ventilated patients in the sample and assuming imperfect serial correlation, suggests that the average number of beyond- \bar{T} ventilation patients is small, and supports the modeling choice of \bar{T} . (A larger $\hat{\alpha}$ would suggest that lagged I_{t-T} variables with $T > \bar{T}$ should be added to improve the model's accuracy.)

Does the model make sense *outside* the sampling period? The necessary assumption for the model to hold outside the sampling period is the following.

Conjecture 1. *The probabilities $\hat{\pi}_T$ ($0 \leq T \leq \bar{T}$) estimated inside the sampling period are relevant outside the sampling period.*

Naturally, there are many examples that fail this assumption. They are discussed in the next section. However, a good performance of the model inside a long sampling period, according to the above criteria, would be encouraging for out-of-sample predictions: it would suggest that the variation of the p_T 's over time was not significant in the past, and that a model of fixed probabilities may not be a bad approximation. Specifically, in the few days just after the sampling period, a well-performing model inside the sample is likely to be good out of sample as well, assuming that changes are unlikely to occur suddenly.

1.3 A more practical version of the model

In the absence of enough data, decreasing the model resolution to avoid overfitting may be necessary. The model can be rewritten using n -size chunks of lagged new confirmed infections, and using a weighted sum of ventilation probabilities when looking forward from these chunks:

$$\begin{aligned}
 E(V_t | \mathbf{I}_t) &= \sum_{T=0}^{\bar{T}} p_T I_{t-T} \\
 &= \sum_{T \in \{0, n, 2n, \dots, \bar{T}-n\}} \bar{p}_{T, T+n-1} \frac{1}{n} (I_{t-T} + I_{t-T-1} + \dots + I_{t-T-n+1}) \\
 &\quad \sum_{T \in \{0, n, 2n, \dots, \bar{T}-n\}} \bar{p}_{T, T+n-1} \left(\frac{1}{n} \sum_{i=0}^{n-1} I_{t-T-i} \right), \tag{4}
 \end{aligned}$$

where

$$\bar{p}_{T, T+n-1} = \sum_{i=0}^{n-1} \frac{I_{t-T-i}}{\frac{1}{n} \sum_{i=0}^{n-1} I_{t-T-i}} p_{T-i}$$

is a weighted sum of ventilation probabilities. Based on this more practical version we can estimate a regression model similar to 3, in which the explanatory variables are *averages* of lagged confirmed infection rates in chunks of n days,

$$V_t = \alpha + \sum_{T \in \{0, n, 2n, \dots, \bar{T}-n\}} \tilde{\pi}_{T, T+n-1} \left(\frac{1}{n} \sum_{i=0}^{n-1} I_{t-T-i} \right) + e_t.$$

The model is interpreted similarly to before.

2 Prediction Caveats

The next examples illustrate how a model of fixed p_T probabilities may leave out important parts of reality, diminishing its explanatory and predictive power.

The end of an outbreak (a mechanical caveat). A mechanical population change emerges as we approach the end of an outbreak: since there are few new confirmed infections, the sample of ventilation patients consists increasingly of patients who are more likely to remain ventilated long after being confirmed. A positive constant is expected to be estimated to account for these ventilation numbers, and the model's explanatory power diminishes.

Changes in the virus’s behavior. The virus might change its behavior over time. For example, it might lead to more severe outcomes in the winter than in the summer. Suppose for simplicity that people are confirmed as infected both in the winter and the summer in similar rates, but that ventilation is needed much more during winter. In that case *extrapolation* between seasons is impossible—the model will overestimate the out-of-sample number of ventilation cases in the summer based on a winter’s in-sample data, and underestimate this number in the winter based on summer data. Even if one has data for an entire year, *in-sample* predictions will be biased by season despite being unbiased on average—predictions for either the summer or winter will be the average of the two seasons, missing the true values. Technically, the model is misspecified: instead of including a single set of ventilation probabilities $\{p_T\}_{T=0}^{\bar{T}}$, it should at least include two sets for the different seasons $\{p_{\text{summer},T}\}_{T=0}^{\bar{T}}, \{p_{\text{winter},T}\}_{T=0}^{\bar{T}}$.

Changes in the infected population. The population attacked by the virus might change over time. For example, a first wave of people confirmed as infected may include mainly young and healthy people who just returned from a ski vacation in Italy, while a second wave may include mainly old people in nursing homes, where the virus is able to rapidly infect many people. Suppose for simplicity that in the beginning of the outbreak only young and healthy people are infected, while in the end of the outbreak only old and ill people are infected. In that case *extrapolation* from a the young-period to the old-period is impossible—it will underestimate the old people’s outcomes. And again, even if both periods of the outbreak are covered by the data, *in-sample* predictions will be off despite being unbiased on average. The model is in that case again misspecified: instead of including on a single set of ventilation probabilities $\{p_T\}_{T=0}^{\bar{T}}$, it should include at least two sets for the different populations $\{p_{\text{young},T}\}_{T=0}^{\bar{T}}, \{p_{\text{old},T}\}_{T=0}^{\bar{T}}$.

Changes in testing and diagnosing the population The population diagnosed and confirmed as infected might change over time. For example, in the beginning of the outbreak, when people are still unaware of the virus and its symptoms, it is possible that only people with severe conditions get tested and classified as infected, and those, by nature, have different ventilation probabilities than people with mild symptoms. This might happen in the end of an outbreak as well: when public discipline looses, people with mild symptoms may become less worried about the virus and avoid being tested, tilting the probabilities again. Conversely, if the Health Ministry starts doing random tests and identifying much more people with mild symptoms than before, the probabilities are tilted the other way.

Changes in treatment practices. Doctors may decide, whether as a part of a large-scale policy or as independent judgments, to treat severe-condition patients differently over time. For example, when experiencing a steep increase in patient numbers, ventilation decisions may be more prudent in order to save space for later patients. When the number of patients is decreasing and the outbreak is considered under control, the alternative cost of ventilating one patient is low, and ventilation may last longer to keep patients out of danger.

3 Data and Design

In order to examine the possible connection between newly confirmed patients and future ventilation numbers, we use publicly available data provided by the Israeli Ministry of Health, hereafter MOH. During the outbreak of covid-19 in Israel, the MOH regularly uploads notifications and reports to their Telegram account.¹

Up until May 3, 2020, the MOH provided a daily report with detailed information regarding the confirmed-infected patients, including the number of daily newly confirmed and currently ventilated, starting from the initial date of the covid-19 outbreak. On May 4, 2020 the MOH stopped providing this source, but it still uploads a daily overview, which includes the current total number of confirmed-infected and the current total number of ventilation patients. Using these two sources, we were able to construct 78 observations of confirmed-infected patients which span from March 1, 2020 to May 17, 2020, and 73 observations of confirmed-infected patients which span from March 6, 2020 to May 17, 2020.²

Using these data, we estimate a set of linear regressions based on the more practical version of the model, in which “Ventilated” at time period t is the dependent variable and the average newly confirmed cases between two previous dates serve as a set of independent variables. For example, the first variable in the following Table 1, “Day 0 to Day -4” is the average number of newly confirmed cases between time periods t and $t - 4$.

4 Estimation Results for Israel: March 6 – May 17

The results of the linear regressions are described in Table 1.

We chose $\bar{T} = 19$ days and $n = 5$ days mostly to keep the degrees of freedom limited in light of the relatively small dataset. As more data accumulates, these choices should be

¹<https://t.me/MOHreport>

²We are unsure whether there were zero ventilation patients or a positive number of ventilation patients in the first five days from March 1 to March 6, in which data on ventilation patients is missing. In all these days there were no more than 3 new confirmed patients per day, and the first published number of ventilation patients in March 6 is 1. Hence any omitted values are of small weight compared to the rest of the sample.

Table 1: Regressions Table

Newly Confirmed During:	(1)	(2)	(3)	(4)	(5)
Day 0 to Day -4	0.15 (0.04)				0.03 (0.01)
Day -5 to Day -9		0.19 (0.03)			0.06 (0.02)
Day -10 to Day -14			0.18 (0.02)		0.07 (0.01)
Day -15 to Day -19				0.13 (0.02)	0.06 (0.01)
Constant	33.84 (19.26)	26.45 (13.49)	29.79 (7.60)	48.59 (12.14)	25.16 (6.98)
Observations	73	69	64	59	59
R^2	0.44	0.78	0.87	0.54	0.95
Adjusted R^2	0.43	0.77	0.87	0.53	0.95

Notes: OLS regressions. Dependent variable: ventilated at time period t . Independent variables: average daily confirmed cases in 5-day periods. Standard errors (in parenthesis) are adjusted using the Newey-West method for addressing autocorrelation and heteroskedasticity, with an autocorrelation parameter of maximal lag defined as 19 days.

revisited. The full model in column (5) is estimated with an $R^2 = 95\%$, which indicates that the dynamics of ventilation numbers in the existing sampling period is explained extremely well with our four regressors. $\hat{\alpha} = 25$ is small relative to the typical numbers of ventilated people (period average = 67) and the $\tilde{\pi}_{T, T+n-1}$ coefficients are all positive. These estimates suggest that a simple model of fixed probabilities does well within the sampling period we use. As discussed above, such a good performance of the model over this period is encouraging regarding its predictive power for future outcomes but—as also discussed above—any extrapolation would crucially rely on assumptions that only future data could confirm or reject. Caution is therefore required.

The figure in the main text shows that the model systematically overestimates ventilation numbers in early dates, and systematically underestimates them in later dates. This suggests a systematic end-of-sample change of populations as described in the previous section, and perhaps other mechanisms that change the ventilation probabilities over time.