

## מודל פשוט למעקב אחרי מספר החולים במצב קשה בישראל

אורי חפץ, עופר גליקסון, וגיא ישי<sup>1</sup>

5 ביולי, 2020

### השאלה

הנחה מרכזית במאבק בנגיף הקורונה היא שיש להימנע מהמצב הנורא בו חולה במצב קשה שזקוק לטיפול רפואי לא יוכל לקבלו. למערכת הבריאות יכולת מוגבלת לטפל בחולים במצב קשה. **כיצד נוכל להתריע מראש על התקרבות למצב שבו המערכת בשיא תפוסתה?**

### המודל

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

ב-17 במאי פרסמנו גרסה ראשונה של מאמר זה בה בחנו את התאמת המודל לנתונים העדכניים דאז, וקיבלנו התאמה מצוינת של המודל לנתונים: מספר חולי הקורונה הקשים

---

<sup>1</sup> האוניברסיטה העברית בירושלים

בישראל ביום נתון ניתן היה לחיזוי כמעט מושלם על סמך מספר חולי הקורונה המאומתים החדשים מהשבועות שקדמו לו.<sup>2</sup> הגרף להלן ממחיש זאת בחלקו השמאלי, עד ה-17 במאי.



העקומה האדומה מראה את כמות הנדבקים המאומתים החדשים בכל יום, החל מה-1 במרץ 2020 ועד ה-4 ביולי והכחולה מראה את מספר החולים במצב קשה (הכולל) בכל יום, החל מה-1 במרץ 2020 ועד ה-4 ביולי, על פי נתוני משרד הבריאות. העקומה התכולה מראה את תחזית המודל לגבי מספר החולים היומי במצב קשה. התחזית מבוססת על רגרסיה ליניארית של מספר החולים היומי במצב קשה על ארבעה ממוצעים של מספר הנדבקים החדשים ב-20 הימים האחרונים. הנוסחה המתקבלת עבור נתוני ה-17 במאי, שמופיעים בחלקו השמאלי של הגרף עם הרקע הלבן, פשוטה מאוד (פרטים בנספח):

$$\text{מספר החולים במצב קשה ביום נתון} = 34 \text{ (מספר קבוע)}$$

- + 2% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים האחרונים (כולל היום)
- + 7% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים שלפניהם (לפני 5-9 ימים)
- + 12% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים שלפניהם (לפני 10-14 ימים)
- + 8% ממספר הנדבקים המאומתים החדשים הממוצע בחמשת הימים שלפניהם (לפני 15-19 ימים).

מה אומרות המשקלות בנוסחה? נניח שבכל יום 100 נדבקים מאומתים חדשים. כעבור 20 יום, המודל חוזה 28 חולים קשים (2+7+12+8, פחות 1 בשל עיגולים) מעבר למספר הקבוע (34). מה עם הקבוע 34? אנו מראים בנספח התיאורטי שהוא מרמז על טעויות במודל (למשל, יש חולים קשה שאומתו כנדבקים לפני 20 הימים האחרונים) או בנתונים.

<sup>2</sup> ראה: <http://users.nber.org/~heffetz/papers/ventilators-as-a-function-of-reported-infected-plus-appendix.pdf>. הגרסה הראשונה התייחסה לכמויות של מונשמים ולא של חולים במצב קשה, אך ההתאמות דומות מאוד בשני המקרים.

## 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?

**בחלקו הימני של הגרף, החל מה-18 במאי**, אנו מנסים לתת תשובה מסוימת לשאלה זו. הנוסחה שהתקבלה על פי נתוני ה-17 במאי משמשת כדי לנסות לחזות את כמויות החולים במצב קשה בתקופה שלאחר מכן. ניתן לראות שהנוסחה נותנת חיזוי יתר שיטתי ביחס לנתוני האמת. למשל, החל מה-29 ביוני ועד ה-4 ביולי הנוסחה חוזה עלייה בכמות החולים במצב קשה מכ-100 לכ-150 בעוד שהעלייה בפועל היא מכ-40 לכ-80. המסקנה היא שהמתאם הסטטיסטי שהתקיים עד ה-17 במאי השתנה בצורה ניכרת לאחר מכן. מה השתנה? מסתבר שאוכלוסיית הנדבקים אינה יציבה לאורך זמן—נתונים שנפרט בהמשך מראים שהיא הופכת צעירה יותר במוצע בגל השני של ההדבקות—אך יכול להיות שגם שינויים במדיניות הבדיקות ובמדיניות האשפוזים עומדים מאחורי הפער שנוצר בין התחזית לנתוני האמת.

האם עדיין יש ערך למודל? כל עוד לא יחולו שינויים חדים בהרכבי האוכלוסייה הנדבקת, ניתן להשתמש בנוסחה כחסם עליון שמרני למדי לכמויות החולים במצב קשה שיש בפועל. בנוסף, ניתן ללמוד משמרנותו של המודל שמבחינת הסיכוי שנדבק חדש יהפוך לחולה קשה, תקופת הקורונה של סוף מאי ויוני הייתה מסוכנת פחות מהתקופה שלפניה.

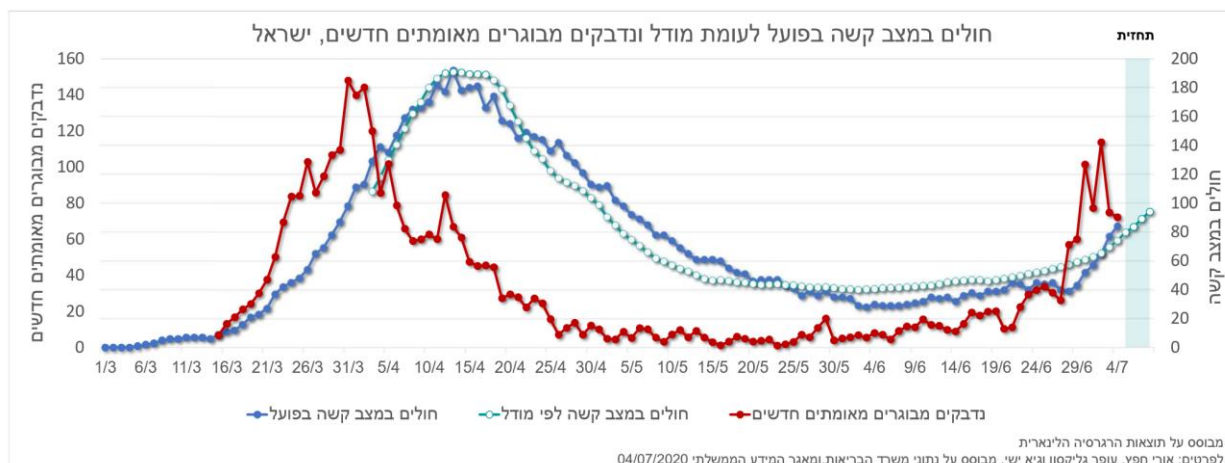
## שיפור המודל על ידי התחשבות בגיל החולים

אי הצלחתו של המודל בחיזוי העתיד קשורה לכך שמאפייני האוכלוסייה הנדבקת והנדבקת, ואולי גם השפעות הנגיף ואופני הטיפול בו, אינם קבועים אלא משתנים באופן שיטתי לאורך זמן. מאפיין מרכזי שעשוי להשתנות עם הזמן הוא שיעור הנדבקים המבוגרים, שסיכוייהם להיזדרזר למצב קשה גבוהים בהרבה משל נדבקים צעירים.

בחלק זה אנו בוחנים מודל שקושר את מספר חולי הקורונה הקשים ביום מסוים למספר חולי הקורונה החדשים **המבוגרים, שמוגדרים כבני 60 לפחות**, שאומתו בימים שקדמו לו. יש לשים לב: המודל עדיין משמש לניבוי הכמות הכוללת של החולים במצב קשה, כולל הצעירים שבהם. ההיסמכות על מספר החולים המבוגרים החדשים נועדה לשפר את הקשר הסטטיסטי שעומד ביסוד המודל, ולהיפטר מ"רעש" שקשור לכמויות משתנות של נדבקים צעירים, שבפועל כמעט ולא מגיעים למצב קשה. הקשר הסטטיסטי ישתפר במידה והקשר בין כמויות נדבקים **מבוגרים** חדשים לבין כמויות חולים במצב קשה (בכל גיל) יוכח כעקבי ויציב יותר לאורך זמן מהקשר שניבחן במודל המקורי, בין כמויות נדבקים חדשים **בכל גיל** לבין כמויות חולים במצב קשה (בכל גיל).

הנתונים אודות חולים חדשים בני 60 לפחות זמינים עבור תקופה קצרה יותר מהנתונים אודות חולים מאומתים חדשים בכל הגילאים, ומטעמי שמירה על פרטיות החולים הנתונים אינם

מדויקים, בייחוד בימים בהם היו מעט חולים מאומתים חדשים. לכן בחלק זה אנו נאלצים ליצור אומדן גס לכמות החולים המבוגרים המאומתים החדשים בכל יום נתון (פרטים בנספח), ומסתפקים באמידת המודל החדש על בסיס כלל הנתונים הזמינים, עד ה-4 ביולי. הגרף להלן ממחיש את ביצועי המודל החדש.



המודל מצליח להסביר בצורה סבירה את כלל הנתונים הזמינים. התוצאות מחזקות את ההשערה שהתחלואה הקשה נותרה נמוכה בתחילת גל הקורונה השני למרות גידול חד בכמויות הנדבקים החדשים, כיוון שכמות הנדבקים המבוגרים המאומתים החדשים נותרה נמוכה. נוסחת הרגרסיה הליניארית שמתקבלת במקרה זה היא:

$$\text{מספר החולים במצב קשה ביום נתון} = 34 (\text{מספר קבוע})$$

- + 6% ממספר הנדבקים המבוגרים החדשים הממוצע בחמשת הימים האחרונים (כולל היום)
- + 30% ממספר הנדבקים המבוגרים החדשים הממוצע בחמשת הימים שלפניהם (לפני 5-9 ימים)
- + 49% ממספר הנדבקים המבוגרים החדשים הממוצע בחמשת הימים שלפניהם (לפני 10-14 ימים)
- + 68% ממספר הנדבקים המבוגרים החדשים הממוצע בחמשת הימים שלפניהם (לפני 15-19 ימים).

תחת ההנחה שהקשר הסטטיסטי שעומד ביסוד המודל יישאר יציב יותר לאורך זמן מהקשר שנידון קודם לכן, ניתן לנסות לחזות את העתיד בעזרתנו בזהירות הראויה. הגרף לעיל מראה את מספר החולים במצב קשה החזוי בארבעת הימים שאחרי אמידת המודל. (מכיוון שאיננו יודעים את מספר הנדבקים המאומתים המבוגרים החדשים בארבעת הימים הבאים, החיזוי משמית 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. ניתן לחשב, למשל, שעם 100 נדבקים מבוגרים חדשים ביום, המודל חוזה 187 חולים במצב קשה החל מהיום ה-20 (6+30+49+68 ועוד הקבוע 34).

אפשר לשאול שאלות מעניינות יותר, כמו: כמה חולים במצב קשה המודל חוזה ביום ה-20 אם נתחיל עם 50 נדבקים מבוגרים חדשים ביום הראשון, ונגדיל את המספר ב-10% בכל יום? אלו מספרים שנראים רלוונטיים נכון לשבועיים האחרונים. במקרה כזה סך הנדבקים המבוגרים החדשים ביום ה-20 יגיע ל-306, והמודל מנבא שכמות החולים במצב קשה ביום זה תהיה 187. הטבלה להלן מחשבת את חיזויי המודל תחת 25 תרחישים כאלו. בפינה השמאלית התחתונה שלה, הטבלה מראה שאם ביום הראשון יש 125 נדבקים מבוגרים חדשים, ואם המספר יגדל בקצב מהיר של 20% ביום, ביום ה-20 יהיו בישראל 907 חולים במצב קשה. זו כבר כמות משמעותית של חולים במצב קשה שממחישה את החשיבות שבשליטה בהתפשטות הנוכחית.

### טבלת ניתוח רגישות למספר החולים במצב קשה החזוי ביום 20

שיעור הגידול היומי של נדבקים מבוגרים מאומתים חדשים						
20%	15%	10%	5%	0%		
208	147	110	87	72	25	מספר הנדבקים המבוגרים המאומתים החדשים ביום 1
383	261	187	140	110	50	
558	375	263	193	149	75	
732	489	340	246	187	100	
907	603	416	299	225	125	

צריך כמובן להיזהר בחיזוי: לא היו בישראל עדיין ימים עם יותר מכ-150 נדבקים מבוגרים מאומתים חדשים, ולא היו בישראל רצפים של 20 יום שמתחילים ב-125 מבוגרים מאומתים חדשים וממשיכים עם קצב גידול של 20%. אם הרכב האוכלוסייה הנדבקת או הנבדקת בתרחישים כאלה יהיה שונה מהותית מזה שהיה בישראל בשלושת החודשים האחרונים, גם המודל הזה 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.<sup>3</sup> הם לוקחים בחשבון נתונים נוספים, למשל פילוח מלא לפי גיל, מגדר ומחלות רקע. 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 במודלים הפשוטים שהצענו במאמר זה, ושהבנתם תעזור לתת לציבור תחושת שליטה, הבנה, וביטחון.

<sup>3</sup> [מספר מספר 147](#) שפורסם ע"י מרכז המידע והידע הלאומי למערכה בקורונה, משתמש בנתונים עשירים מאלה שזמינים עבורנו ומגיע למסקנות דומות אודות התפתחות המחלה באוכלוסייה המבוגרת.

## סיכום

בגל הקורונה הראשון, עד ה-17 במאי, ניתן היה להסביר את מספר החולים במצב קשה בכל יום כפונקציה פשוטה של מספר הנדבקים החדשים בשלושת השבועות שלפניו. הסבר זה מספק תחזית לא מדויקת לגל הקורונה השני, וניתן להתייחס לתחזית שלו כחסם עליון שמרני לתחלואה הקשה בתקופה זו ובעתיד הקרוב. שימוש במספר הנדבקים **המבוגרים** החדשים עשוי להסביר את כמות החולים במצב קשה באופן מספק יותר. בתחילתו של גל הקורונה השני נדבקו פחות מבוגרים מאשר בגל הראשון, מה שעשוי להסביר את מיעוט מקרי התחלואה הקשה עד לסוף יוני. יחד עם זאת, לאחרונה נצפית עלייה בכמות הנדבקים המבוגרים החדשים ובתחלואה הקשה. אם הקשר הסטטיסטי מהחודשים האחרונים יימשך, נוכל לדעת מבעוד מועד אם אנחנו מתקרבים לעומס יתר על מיטות האשפוז ומכונות ההנשמה. כרגע נראה שאנחנו עוד רחוקים, אך **התוצאות שבידינו ממחישות את החשיבות שבמניעת הדבקת האוכלוסיה המבוגרת וקבוצות הסיכון האחרות.**

# A Model of Covid-19 Critical-Condition Patient Numbers - Appendix

Ofer Glicksohn

Ori Heffetz

Guy Ishai

July 5, 2020

## 1 The Model

Our goal is to explain and predict the number of critical-condition 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 critical patients in day  $t \in \mathcal{T}$  by  $V_t$ . Becoming critically ill 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 be in a critical condition 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 becomes critical or not within a limited time interval, and those who become critical 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 the critical-condition probabilities, the number of newly confirmed patients in day  $t - T$  who are critical in day  $t$ , i.e., the *contribution* of day  $t - T$  to the number of critical 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 critical 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 critical 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 being in a critical condition 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 critical in day  $t$  by  $p_\infty$ . Therefore

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

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:

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

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  $I_{t,t-T}$  variable for  $0 < T \leq \bar{T}$ , such that the following linear projection can be estimated based on the sample:

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

where  $E(\nu_t \mathbf{I}_t) = \mathbf{0}$ . The constant  $\beta$  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  $\beta$  is just the sample mean of  $I_{t,\infty}$ . Substituting in equation 2 we derive

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

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

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

provides consistent estimates for  $\alpha$  and  $\pi_T$ .

The estimated  $\hat{\pi}_T$ 's can no longer be interpreted as proxies of the critical-condition prob-

abilities  $p_T$  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  $I_\infty$ , and on the aggregate probability of the omitted confirmed infections to be critical patients at day  $t$ ,  $p_\infty$ . 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 are critical beyond  $\bar{T}$  days after confirmation, which is *not already explained* by the  $I_{t-T}$  variables based on their correlation with  $I_\infty$ . 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 critical patients in the sample and assuming imperfect serial correlation, suggests that the average number of beyond- $\bar{T}$  critical 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 critical-condition 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 critical-condition 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 critical patients consists increasingly of patients who are more likely to remain critical long after being confirmed. A positive constant is expected to be estimated to account for these critical patients 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 critical condition is more likely during winter. In that case *extrapolation* between seasons is impossible—the model will overestimate the out-of-sample number of critical 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 critical-condition 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 critical-condition 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 dif-

ferent critical-condition 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, decisions about critical condition 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 defining one patient as critical is low, and critical situation 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 critical cases, we use publicly available data provided by the Israeli Ministry of Health, hereafter MOH. The MOH website updates a historical dataset on a daily basis.<sup>1</sup> The dataset contains complete information of the daily cumulative number of confirmed patients and the daily number of critical patients, from the beginning of the covid-19 outbreak (January 26, 2020) up to a week before the current date. The dataset we use was retrieved on July 5, 2020 and contained complete data up to June 28, 2020. Since almost all observations prior to March 1, 2020 contained no newly confirmed cases and no patients in critical condition, and in order to be consistent with our results from May 17, 2020, we omitted those observations.

The MOH also maintains an interactive dashboard of the current overview for the covid-19 situation.<sup>2</sup> This dashboard is updated 3 times a day and contains the number of newly confirmed patients for each day in the past month. The dashboard is used to get the number of newly confirmed patients in the dates between June 29, 2020 up to July 4, 2020, which are not covered by the historical dataset. However, the dashboard does not include full data about the number of critical patients in those dates. To get these numbers, we use recent snapshots of the dashboard, which are regularly uploaded by the MOH to their Telegram account.<sup>34</sup>

---

<sup>1</sup><https://govextra.gov.il/ministry-of-health/corona/corona-virus/>

<sup>2</sup>[https://datadashboard.health.gov.il/COVID-19/?utm\\_source=go.gov.il&utm\\_medium=referral](https://datadashboard.health.gov.il/COVID-19/?utm_source=go.gov.il&utm_medium=referral)

<sup>3</sup><https://t.me/MOHreport>

<sup>4</sup>Comparing the dashboard to the historical data reveals that the dashboard is mostly accurate towards

Using these sources we were able to construct 126 observations of critical patients and newly confirmed patients, which span from March 1, 2020 to July 4, 2020.

## Demographic Data

We later also examine the possible connection between the number of newly confirmed *elderly* patients and the future number of critical-condition patients. We define elderly as being above 60. We use another dataset publicly provided by the MOH, which specifies the *weekly* number of observations in each combination of gender (Male, Female, Null) and age group (0-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+, Null).<sup>5</sup> The dataset spans from March 15, 2020 to June 27, 2020. Since the age and gender of a newly confirmed patient is considered private, there is an important caveat in this dataset: when the number of observation in a given (gender, age) combination is between 1 and 14 inclusive, the only information provided is the number being “< 15”. Due to these limitations of a weekly resolution and missing numbers, we can only create a rough estimate for the daily share of the elderly out of the newly confirmed . We do so by applying the following procedure:

1. In any (gender, age) combination in which the provided number is “< 15”, and in which gender is “Null”, we assign the number 0. We do so as we suspect that the actual number is close to zero.
2. In any (gender, age) combination in which the provided number is “< 15”, and in which gender is either “Male” or “Female”, we assign the number 4. The number 4 is the constant number for which the sum of newly confirmed cases, summed across all combinations and across all this dataset (March 15, 2020 to June 27, 2020), resembles the comparable number of total newly confirmed patients in the historical dataset provided by MOH.
3. We assume that the number of newly confirmed patients above 60 in the “Null” groups is zero.
4. We calculate the weekly share of the number of elderly out total number of all (gender, age) combinations, and apply it for each day of a given week.

This procedure creates a rough estimate for true share of the elderly out of the newly confirmed patients. Most of the missing data is from May, hence our numbers of newly confirmed

---

the end of the day. Hence used the most recent daily snapshot for each day. Each day’s most recent snapshot is always from later than 17:00.

<sup>5</sup><https://data.gov.il/dataset/covid-19/resource/89f61e3a-4866-4bbf-bcc1-9734e5fee58e?inner>

elderly patients for this month are less accurate than the rest of the data. However, this is closely related to the fact that there were small overall numbers of new infections in May. Therefore the effect that the missing data has on predicted numbers of critical-condition patients is limited.

To complete the dataset up to July 4, 2020, we also use a daily report from July 5, 2020 published in the covid-19 national info center website<sup>6</sup>. The report includes a cumulative bar diagram for the daily distribution of newly confirmed patients by age groups (0-19, 20-39, 40-59, 60-79, 80+, Unknown) spanning from June 20, 2020 to July 4, 2020. This bar diagram shows the share of the age groups (0-19, 20-39, 40-59) among the newly confirmed patients, which are used to complete the dataset in the dates June 28, 2020 to July 4, 2020. A technical difficulty is that the shares of “Unknown” patients are not detailed inside the diagrams. These are small shares relative to the entire diagram and we assign the constant number of 3% to represent them, as the share of the “Null” groups in the weekly dataset above, for the last available week, is also roughly 3%.

By combining these two data sources we have 112 daily estimates for the share of the elderly out of the newly confirmed patients, spanning from March 15, 2020 to July 4, 2020. We use this estimate, combined with the daily number of newly confirmed patients, to create the estimated daily number of newly confirmed elderly patients.

## 4 Estimation Results for Israel

### 4.1 Basic Model Estimation: March 1 – May 17

In order to be consistent with our work from May 17, 2020, we use data up to May 17, 2020 (78 observations), to estimate a set of linear regressions based on the more practical version of the model, in which “Critical” 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. The results are described in Table 1. 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  $t - 4$ .

We chose  $\bar{T} = 19$  days and  $n = 5$  days mostly to keep the degrees of freedom limited. The full model in column (5) is estimated with an  $R^2 = 99\%$ , which indicates that the dynamics of critical-condition numbers in the existing sampling period is explained extremely well with our four regressors. The good fit suggests that a simple model of fixed probabilities does well within the sampling period we use. As discussed above, such a good performance of

---

<sup>6</sup><https://www.gov.il/he/departments/publications/reports/daily-report-05072020>

Table 1: Basic Model Estimation: March 1 to May 17

Newly Confirmed During:	(1)	(2)	(3)	(4)	(5)
Day 0 to Day -4	0.19 (0.06)				0.02 (0.01)
Day -5 to Day -9		0.25 (0.04)			0.07 (0.01)
Day -10 to Day -14			0.26 (0.02)		0.12 (0.01)
Day -15 to Day -19				0.19 (0.02)	0.08 (0.01)
Constant	47.82 (24.00)	37.84 (15.78)	38.94 (6.99)	60.64 (15.31)	33.73 (4.53)
Observations	74	69	64	59	59
$R^2$	0.41	0.76	0.94	0.62	0.99
Adjusted $R^2$	0.40	0.75	0.93	0.62	0.98

**Notes:** OLS regressions. Dependent variable: critical patients 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.

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 (see next subsection).

The first figure in the main text up to May 17 visually shows the fit of the model.

## 4.2 Prediction: May 18 – July 4

To test the predictive power of the model, we try to predict critical-condition numbers after May 17, 2020 using the estimation results from above. The first figure in the main text shows the goodness of fit of the model in the period beginning in May 18, 2020. The model fails to predict this additional period, and systematically overestimates the number of critical patients by a factor which grows to approximately 2 towards the end of the period. This may suggest that the critical-condition probabilities systematically decrease between the first wave of covid-19 spread and the second wave. Possible reasons for such a change were discussed in Section 2. We suggest that the model’s predictions can still be used as an upper-bound of true critical-condition outcomes, under the assumption that the underlying probabilities indeed systematically decreased.



### 4.3 Re-estimating the Model: March 1 – July 4

The previous results suggest that a model of fixed critical-condition probabilities cannot account for all the variation in the data, a concern raised in Section 2. To test the extent to which this concern is true, we re-estimate the model based on the full dataset we have. The results are described in Table 2 and the next figure.

Table 2: Second Model Estimation: March 1 to July 4

Newly Confirmed During:	(1)	(2)	(3)	(4)	(5)
Day 0 to Day -4	0.12 (0.06)				-0.07 (0.04)
Day -5 to Day -9		0.22 (0.04)			0.13 (0.04)
Day -10 to Day -14			0.26 (0.01)		0.12 (0.04)
Day -15 to Day -19				0.23 (0.02)	0.07 (0.02)
Constant	44.80 (16.65)	28.67 (10.25)	26.55 (6.96)	36.23 (10.76)	29.05 (5.91)
Observations	122	117	112	107	107
$R^2$	0.22	0.62	0.85	0.71	0.88
Adjusted $R^2$	0.22	0.62	0.85	0.71	0.87

**Notes:** OLS regressions. Dependent variable: critical patients 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.



We find that there is no significant improvement in the overall goodness of fit, especially in the overestimated period of June. Importantly, the fit during the second wave does slightly improves, but on the expense of the fit in first wave, which is now worse than before. This supports the hypothesis that the critical-condition probabilities systematically change

between the two covid-19 waves, making a model of fixed probabilities irrelevant from both waves together, as well as for predicting one wave based on the other (as detailed in Section 2).

#### 4.4 Estimating a Model Based on Age of Newly Confirmed: March 15 – July 4

To have a better fit in the entire estimation period and in order to make better predictions, we estimate a new and different model from the main one described in the paper. The new model relies on quantities of newly confirmed *elderly* patients only, rather than on total quantities of newly confirmed patients. The model relies on the empirical finding that a large portion of critical-condition patients are aged 60 and above, with the average age of such patients being 65-70.<sup>7</sup> Hence the number of newly confirmed patients aged below 60 provides a noisy signal for critical-condition outcomes, and a model based on newly confirmed elderly patients alone only can potentially provide better results. A caveat when estimating this new model is that the interpretation of coefficients gets further away from the original idea of critical-condition *probabilities*, due to the fact that not every critical patient is aged 60 or above.

We estimate the model using the full period of March 15, 2020 to July 4, 2020 rather than using the first period which ends at May 17, 2020, since the age-specific data only begins in mid March and the data has a lower quality than before. The results are described in Table 3.

The model's fit over the entire period is better than the previous model which does not account for the age of the newly confirmed patients. This supports the hypothesis that there were only few critical-condition patients in the beginning of the second covid-19 despite large numbers of newly confirmed patients, because only a small share of those patients are elderly.

---

<sup>7</sup><https://data.gov.il/dataset/covid-19/resource/e4bf0ab8-ec88-4f9b-8669-f2cc78273edd>

Table 3: Elderly-Newly-Confirmed-Based Model Estimation: March 15 to July 4

Newly Confirmed During:	(1)	(2)	(3)	(4)	(5)
Day 0 to Day -4	0.71 (0.24)				0.06 (0.11)
Day -5 to Day -9		1.17 (0.20)			0.30 (0.10)
Day -10 to Day -14			1.35 (0.12)		0.49 (0.16)
Day -15 to Day -19				1.29 (0.13)	0.68 (0.14)
Constant	55.16 (16.48)	44.52 (11.51)	40.06 (7.64)	40.19 (7.74)	33.62 (6.52)
Observations	108	103	98	93	93
$R^2$	0.25	0.66	0.89	0.81	0.95
Adjusted $R^2$	0.24	0.65	0.89	0.81	0.94

**Notes:** OLS regressions. Dependent variable: critical patients at time period  $t$ . Independent variables: average daily estimated confirmed cases of people aged 60 and above 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.