Universität Konstanz

22.4.2020

Schätzung der Reproduktionsrate R anhand der Neuinfektionen nach Meldedatum der SARS-CoV-2 Epidemie in Deutschland

Prof. Dr. Christian Wienbruch

christian.wienbruch@uni-konstanz.de

Universität Konstanz

Klinische Psychologie

PF905

78467 Konstanz

1. Die Reproduktionsrate R[1]

Ein entscheidender Parameter bei der Beurteilung der Situation während der andauernden Corona-Virus-Epidemie (COVID19) durch das SARS-CoV2-Virus ist die Reproduktionszahl R. Das Robert-Koch-Institut (RKI) definiert R als „*die Anzahl der Personen, die im Durchschnitt von einem Indexfall angesteckt werden*“ [1]. Dieser Parameter ist deswegen bedeutsam, weil er den gegenwärtigen Zustand der Epidemie beschreibt und vorsichtige, in die Zukunft gerichtete Prognosen erlaubt. Außerdem kann er dazu genutzt werden, der Allgemeinheit leicht verständlich zu vermitteln, ob die Epidemie kontrolliert oder unkontrolliert verläuft. Ist dieser Wert kleiner als eins, so verläuft die Epidemie kontrolliert und kann gestoppt werden. Ist er größer als 1, so verläuft sie unkontrolliert und kann sich wieder verstärken. Aus diesem Grund ist es wichtig, R möglichst genau und möglichst gegenwartsnah zu berechnen.

R ist unter anderem eine Funktion der Zeit und kann regional variieren. Wird R aus der Zahl der Neuinfektionen geschätzt, so gilt dieser Parameter für die Grundgesamtheit aller Menschen in Deutschland – die Gesamtpopulation - nur dann, wenn die Stichprobe, die auf SARS-CoV2-Viren getestet wird, zufällig bestimmt wurde. Da in Deutschland aber nur eingeschränkt und vor allem selektiv getestet wird, kann ein so geschätztes R nicht als das R der Gesamtpopulation angesehen werden. Derzeit wird nur die Gruppe der Personen, die spezifische Symptome haben oder Kontakt zu Patienten mit COVID19-Erkrankung hatten, getestet. Personen, die mehr oder weniger symptomfrei die Infektion durchlaufen, aber durchaus andere infizieren können, werden nicht oder nur zufällig erfasst. Aufgrund des selektiven Testens können keine Aussagen getroffen werden, ob ein berechnetes R repräsentativ für die Gesamtpopulation ist.

Das RKI beschreibt eine Methode, wie man aus den Meldungen der Neuerkrankungen einen Schätzwert für R bestimmen kann [1]. Die gemeldeten Zahlen werden dabei zunächst nach dem Meldedatum erfasst. Um R korrekt zu berechnen, wäre jedoch die Zahl der Neuinfektionen, erfasst nach dem Infektionsdatum, wünschenswert, was nicht möglich ist, da das Infektionsdatum in der Regel unbekannt ist und somit nur geschätzt werden kann. Als zweitbester Parameter wird von den Autoren der Studie das Erkrankungsdatum genannt, das in 61% aller gemeldeten Fälle angegeben wurde, fehlende Werte werden vom RKI geschätzt[2]. Dieser Prozess wird als *Nowcasting* beschrieben und erlaubt keine Aussagen für die jeweils letzten 3 Tage[3]. Als Ergebnis des *Nowcastings* erhält man die korrigierte Neuinfektionszahl N_c. Das RKI berechnet R anhand dieser Zahl. Die Autoren verweisen jedoch explizit auf den Umstand, dass eine R-Schätzung anhand der täglich gemeldeten Neuinfektionen am

[1] die Reproduktionsrate R ist nicht zu verwechseln mit dem Statistik-Paket R, welches hier zur Analyse verwendet wird

[2] Epid. Bull. 17/2020 S. 11

[3] Epid. Bull. 17/2020 S. 12 - Zitat „*Das Nowcasting verhält sich instabil für Fälle mit einem Erkrankungsbeginn 3 Tage oder weniger vor dem Stand der Analyse, …*“

Meldedatum (N_m) ebenfalls möglich ist[4]. N_m wird vom RKI täglich veröffentlicht[5] und dient als Datenbasis für die folgende Analyse.

2. Die Berechnung von R

Die Autoren vom RKI beschreiben die Berechnung von R so: „*Bei einer konstanten Generationszeit von 4 Tagen, ergibt sich* R *als Quotient der Anzahl von Neuerkrankungen in zwei aufeinander folgenden Zeitabschnitten von jeweils 4 Tagen.*“.[6]

Sei $\mu(i)$ der Mittelwert der Neuerkrankungen der letzten n Tage am Tag i

$$\boldsymbol{\mu(i) = \frac{1}{n} \sum_{j=i-n+1}^{i} N(j)}$$

Dann ist $R(i)$:

$$R(i) = \frac{\mu(i)}{\mu(i-n)}$$

Das Konfidenzintervall für den Mittelwert von R wird aus dem geschätzten Standardfehler des Mittelwerts von R berechnet:

$$\sigma_{\overline{R}} = \frac{\hat{\sigma}_R}{\sqrt{m}}$$

wobei $\hat{\sigma}_R$ die geschätzte Standardabweichung von R und m die Zahl der beobachteten R ist. Das 95%ige Konfidenzintervall ergibt sich:

$$\Delta_{crit} = \pm 1.96 * |\sigma_{\overline{R}}|$$

Für jedes $R(i)$ kann auch ein Konfidenzintervall berechnet werden. Dieses ergibt sich aus dem Standardfehler von $\mu(i)$.

$$\sigma_{\overline{\mu}}(i) = \frac{\hat{\sigma}_{\mu}(i)}{\sqrt{n}}$$

Nach Fehlerfortpflanzung ergibt sich das Konfidenzintervall von $R(i)$ dann zu

$$\Delta_{crit}(i) = \pm 1.96 * \sqrt{\left(\sigma_{\overline{\mu}}(i) / \mu(i-n)\right)^2 + \left(\sigma_{\overline{\mu}}(i-n) * \mu(i) / \mu^2(i-n)\right)^2}$$

[4] Epid. Bull. 17/2020 S. 15

[5] https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Fallzahlen.html

[6] Epid. Bull. 17/2020 S. 15

Wird in den Gleichungen N durch N_c ersetzt und $n = 4$ gewählt, erhält man die RKI-Methode zur Bestimmung von R. Diese wird im Folgenden als NC4-Modell oder kurz NC4 bezeichnet.

Die Zahl der Neuerkrankungen nach Meldedatum N_m stehen seit dem 24. Februar 2020 der Öffentlichkeit zur Verfügung[7]. Wird R aus N_m und n=4 berechnet, so wird das im Folgenden als NM4-Modell oder kurz NM4 bezeichnet. Tests von R (NM4) auf Normalverteilung (Shapiro - Wilk; Tabelle 1) ergeben, dass sowohl für den betrachteten Zeitraum vom 3.3.2020 - 21.4.2020[8] (Abbildung 1A) wie auch 16.3.2020 - 21.4.2020 (Abbildung 1B) keine Normalverteilung vorliegt.

Berechnen wir das Konfidenzintervall von $R(i)$ haben wir im Falle von n=4 eine sehr kleine Stichprobe. Der Standardfehler muss für kleine Stichproben – Normalverteilung des Merkmals (also der $N(j)$) vorausgesetzt, korrigiert werden und zwar mit dem t-Wert der Verteilung [2]. Der Faktor 1.96 ist somit durch $t(\alpha/2 = 0.975, df = 3) = 3.18$ zu ersetzten. Wir haben also durch die Wahl von n=4 eine Situation, in der die Konfidenzintervalle erwartungsgemäß groß sein werden. Das Konfidenzintervall in Abbildung 2 A, C und 3 A, C wurde aus diesen Gründen mit t-Wert Korrektur berechnet. Der vom RKI für den 04-04-2020 angegebene R-Wert[9] und das zugehörige Konfidenzintervall (NC4) stimmt mit den hier berechneten Werten (NM4) überein (Tabelle 2).

$R(i)$ (NM4) ist in Abbildung 2A dargestellt. Das 95% Konfidenzintervall ist blau unterlegt. Die rote Linie markiert $R = 1$, der Wert bei dem die Zahl der Neuerkrankungen konstant bleibt, oberhalb steigt die Zahl der Neuinfektionen, unterhalb ebbt die Pandemie ab.

Statt der klassischen Berechnung des Konfidenzintervalls kann als Alternative z.B. eine lokale Polynom-Regression (LOESS - locally estimated scatterplot smoothing; Local Polynomial Regression Fitting; siehe geom_smooth() aus dem Paket ggplot2 v3.3.0 und die Funktion loess() aus dem Paket stats v3.6.2; [3]) ausgeführt werden (Abbildung 2B). Das 95%-Konfidenzintervall ist auch hier blau unterlegt.

Am 16.März beschlossen Bund und Länder Leitlinien gegen die Ausbreitung des Corona Virus. Um den Einfluss dieser Maßnahmen besser beurteilen zu können, wird ein zweiter Zeitraum betrachtet, der am 16.3.2020 beginnt. Ab diesem Zeitpunkt weist der Verlauf von R schon eine Beruhigung des sprunghaften Verhaltens auf, welches während der ersten 3 Wochen zu beobachten ist (Abbildung 2 C, D).

Die Berechnung von R – als Quotient eines gleitenden Mittelwerts über die Zeit – kann auch als Filter interpretiert werden. Bei der Berechnungsmethode NC4 wird $n = 4$ vom RKI aufgrund der 4-Tages-Generationszeit so gewählt. Das mag sinnvoll sein, um im Mittel alle von einem Indexpatienten neuinfizierten Fälle zu berücksichtigen. Ein Mittelwert über ein Zeitintervall von 4 Tagen ist jedoch ungeeignet, um die wöchentlichen 7-Tage-Variationen im Meldeprozess zu korrigieren. In den letzten 5 Wochen sind solche Schwankungen mit einem Anstieg jeweils zum Wochenende, zu beobachten (Abbildung 2C). Datengeleitet wäre es deshalb besser, $R(i)$ mit $n = 7$ zu bestimmen, also die N_m über einen Zeitraum von 7 Tagen

[7] https://de.wikipedia.org/wiki/COVID-19-Pandemie_in_Deutschland

[8] wird n=4 gewählt so ist liegt der erste R-Wert aufgrund des Schaltjahres am 2.3.2020 vor

[9] Epid. Bull. 17/2020 S. 13

zu mitteln, da dann immer genau ein Wochenende in jedem $R(i)$ berücksichtigt wird[10] und so immer die gesamte, durch die Meldekette verursachte, wöchentliche Varianz enthalten ist. Auch die $R(i)$ (NM7) sind nicht normalverteilt. Tests von R auf Normalverteilung (Shapiro – Wilk; Tabelle 1) ergeben, dass sowohl für den betrachteten Zeitraum vom 8.3.2020 - 21.4.2020[11] wie auch 16.3.2020 - 21.4.2020 keine Normalverteilung vorliegt (siehe auch Abbildung 1C, D). Das Ergebnis der 7-Tage-Mittelung (NM7) ist in Abbildung 3 dargestellt.

Gelöscht: Abbildung 3

Nach NM7 bestimmte $R(i)$ gelten immer 7 Tage retrograd. $R(i)$, nach NC4 bestimmt, sind für die letzten 3 Tage instabil[3]. Folglich beziehen sich diese immer auf einen Zeitraum von 7 bis 4 Tage retrograd. NM7 hat gegenüber NC4 drei Vorteile:

1. $R(i)$ enthält immer die Entwicklung der letzten 3 Tage.
2. das Konfidenzintervall ist kleiner
3. meldekettenbedingte, tägliche Schwankungen sind weitgehend eliminiert

Das beste Modell in diesem Sinn ist NM7 in Kombination mit LOESS. Das Konfidenzintervall dieses Modells ist nochmal kleiner als das des NM7 ohne LOESS. Seit dem 09-04-2020 sind alle $R(i)$ kleiner als 1 und auch das Konfidenzintervall liegt unterhalb oder an dieser wichtigen Marke. Ausgehend von diesem Modell ist die Epidemie in Deutschland seit dem 9. April als kontrolliert zu betrachten. Abgesichert wird diese Aussage mit 95% statistischer Sicherheit, denn auch das Konfidenzintervall liegt unterhalb oder an der roten Linie.

Insgesamt ergeben sich aus dieser Analyse 4 Folgerungen:

1. Die Stichprobe erlaubt keine Aussagen bezüglich der Gesamtpopulation.
2. R kann mit ausreichender Sicherheit aus den Neuinfektionen am Meldedatum (N_m) bestimmt werden.
3. lokale Regressionsmodelle (LOESS) liefern interpretierbare Konfidenzintervalle die stabil über die Zeit sind. Diese Modelle erlauben eine vorsichtige Prognose von R.
4. längere Mittelungsperioden (> 4 Tage) sind aus datengeleiteten Überlegungen heraus geboten. 7-Tage-Mittelungen lösen das Problem der 7-Tage-Variationen.
5. Die SARS-CoV-2 Epidemie ist seit dem 9.4.2020 unter Kontrolle.

Das Stichprobenproblem sollte umgehend behoben werden, damit politische Entscheidungen entsprechend abgesichert getroffen werden können.

[10] in den oben angegebenen Gleichungen ist dann n=7 einzusetzen

[11] wird n=7 gewählt so ist liegt der erste R-Wert aufgrund des Schaltjahres am 8.3.2020 vor

Tabelle 1: Test auf Normalverteilung – Shapiro Wilk

	W	p	Methode
R (02-03-2020 - 21-04-2020)	0.62	<< 0.0001	NM4
R (16-03-2020 - 21-04-2020)	0.93	< 0.02	NM4
R (08-03-2020 - 21-04-2020)	0.87	< 0.0002	NM7
R (16-03-2020 – 21-04-2020)	0.84	<< 0.0001	NM7

Tabelle 2: Konfidenzintervall für den 4.April 2020

Berechnung des Konfidenzintervalls	Datum	R	Konfidenzintervall	Konfidenzintervall
NM4	2020-04-04	1.22	0.81	1.62
NM7	2020-04-04	1.17	0.79	1.54
NC4 [1]	2020-04-04	1.2	0.9	1.6

Statistische Analyse

Die Abbildungen wurden mit ggplot2[12] aus dem R-Paket erstellt [3]. Die nichtlineare Regression der Reproduktionszahl R wurde mit dem Paket geom_smooth[13] und der Methode *loess*[14] (Parameter span=0.4) durchgeführt. Die Datenmatrix, das R-Script, welches die Abbildung 1,2,3 erstellt, sowie die Abbildungen 1,2,3 (werden immer mal wieder aktualisiert) finden sich auf github und kann interessierten Lesern zur Verfügung gestellt werden.

Gelöscht: i

[12] https://www.rdocumentation.org/packages/ggplot2/versions/3.3.0

[13] https://www.rdocumentation.org/packages/ggplot2/versions/3.3.0/topics/geom_smooth

[14] *loess* steht für *locally estimated scatterplot smoothing*

Literatur

1. an der Heiden, M. and O. Hamouda, *Schätzung der aktuellen Entwicklung der SARS-CoV-2- Epidemie in Deutschland – Nowcasting.* Epidemiologisches Bulletin, 2020. **17**: p. 10-15.
2. Bortz, J., *Statistik für Sozialwissenschaftler*. 1999, Berlin Heidelberg New-York: Springer-Verlag.
3. R Core Team, *R: A Language and Environment for Statistical Computing*. 2020, R Foundation for Statistical Computing.

Abbildung 1: Verteilung und Dichte der Reproduktionszahl R

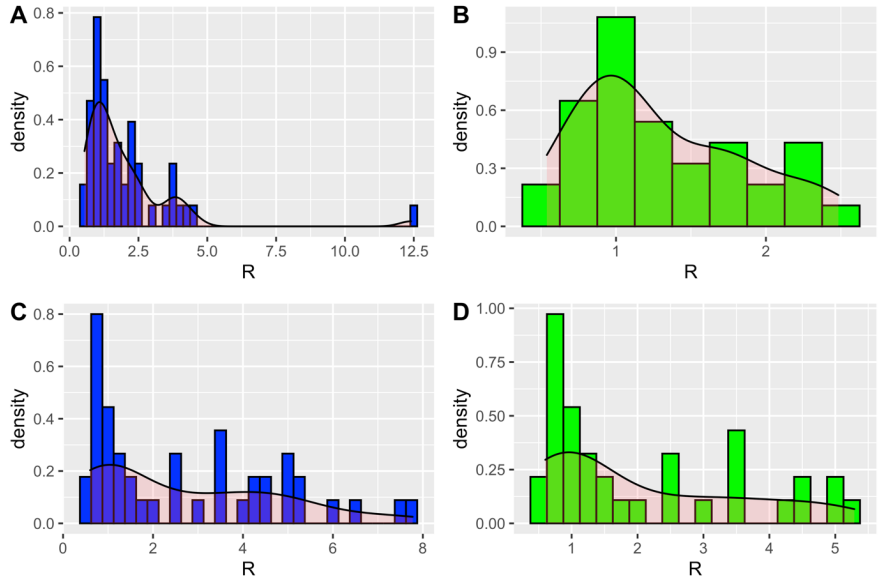

Abbildung 2: Schätzung der Reproduktionszahl R - gleitendes Mittel über 4 Tage

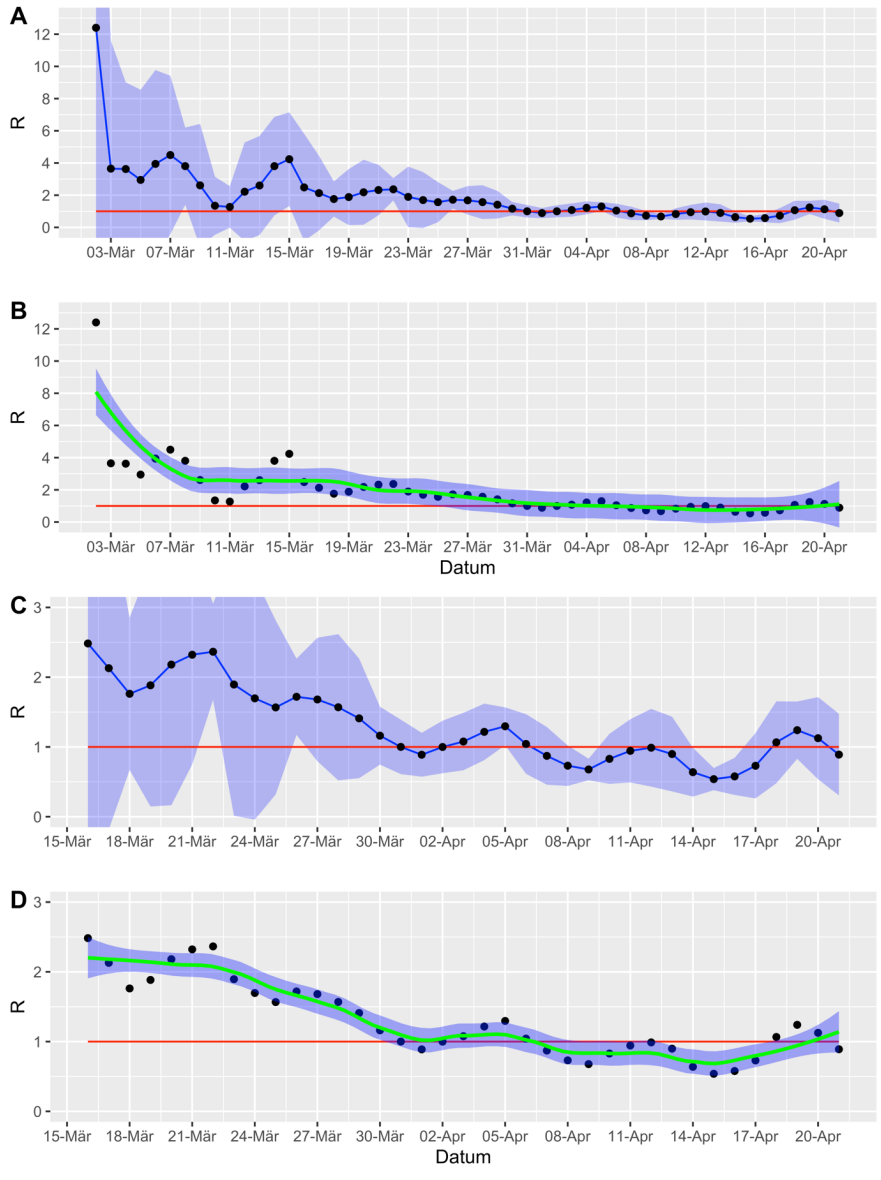

Abbildung 3: Schätzung der Reproduktionszahl R - gleitendes Mittel über 7 Tage