IMPUTATION DES ERKRANKUNGS-DATUMS BEI UNVOLLSTÄNDIGEN DATEN IM RAHMEN DER COVID-19 EPIDEMIE, ÖSTERREICH

Juni 2020

L. Richter, D. Schmid* (Abteilung Infektionsepidemiologie & Surveillance, AGES)
E. Stadlober (Institut für Statistik, Technische Universität Graz)

Österreichische Agentur für Gesundheit und Ernährungssicherheit GmbH (AGES)

Zusammenfassung

Beim Fallbericht von meldepflichtigen Erkrankungen ist häufig das Datum der Erkrankung unvollständig oder verspätet verfügbar, sodass für zeitliche Analysen ein Ersatzdatum, wie zum Beispiel das Labormeldedatum, verwendet werden muss. Wir stellen ein statistisches Modell vor, mit dem die fehlenden Daten für das Erkrankungsdatum basierend auf der berechneten Differenz zwischen Labormeldedatum und Erkrankungsdatum von Fallberichten mit vollständigem Daten geschätzt, sprich imputiert, werden. Das Modell wird auf den COVID19-Surveillance-Datensatz für Österreich angewandt. Die Differenz zwischen Erkrankungsdatum und Labormeldedatum betrug im Mittel 5.4 Tage, mit Variabilität nach Kalenderwoche der Epidemie: die Differenz stieg mit der Fallzahl pro Kalenderwoche. Der Fallzahlgipfel war basierend auf dem Labormeldedatum am 26.03.2020 und bezogen auf das Erkrankungsdatum, inklusive der Fälle mit imputiertem Erkrankungsdatum, bereits am 16.03.2020.

Abstract

The date of disease onset in case reports of notifiable diseases is often missing or reported with delay. Then, a surrogate date, such as the date of the diagnostic laboratory report, has to be used for time series analyses. We introduce a statistical model, which allows the imputation of the missing onset dates based on the known difference between date of lab report and of disease onset, obtained from complete case reports. The model was applied on the Austrian COVID19 surveillance data. The mean difference between lab report date and disease onset was 5.4 days with variation by calendar week of the report: the more case reports per week, the higher the difference. We found the Austrian COVID19 epidemic peaked on March 26, 2020, using lab report dates, compared to an epidemic peak on March 16, 2020, when using known and imputed disease onset dates.

Einleitung

Das Datum der Erkrankung ist eine der wichtigsten Kenngrößen zur Beschreibung und Analyse von Ausbrüchen von Infektionskrankheiten. Allerdings ist üblicherweise die Information über das Erkrankungs-Datum nicht vollständig bzw. nicht zeitgerecht verfügbar. Statistische Methoden erlauben aber die Imputation fehlender Daten basierend auf der bekannten Verzögerung der Fallmeldung.

GÜNTHER et al. (2020) verwenden für Bayern ein *flexible generalized additive model for location, scale and shape* für die Meldeverzögerung in dem für drei Größen kontrolliert wird: der Wochentag der Fallmeldung, die Kalenderwoche der Fallmeldung sowie das Alter der Fälle. Unter Verwendung des R-Pakets `gamlss` können die Parameter von ausgewählten Wahrscheinlichkeitsverteilungen wie der Weibull- oder Gammaverteilung in Abhängigkeit von den Modellgrößen geschätzt werden (RIGBY u. STASINOPOULOS, 2005).

Wir stellen im Folgenden ein ähnliches Modell vor und wenden es auf die österreichischen Surveillancedaten an. Basierend auf der Zeitreihe der vervollständigten Daten zum Erkrankungsbeginn schätzen wir zusätzlich die effektive Reproduktionszahl.

Methoden

Im Folgenden bezeichnen wir die Meldeverzögerung auch mit t_{diff} und sie ist als Differenz von Labormelde-Datum (=Tag der Labormeldung des Falls ins EMS und somit Erfüllung der Falldefinition) und Erkrankungs-Datum definiert.

Imputation des Erkrankungs-Datums

Ähnlich wie GÜNTHER et al. (2020) verwenden wir ein `gamlss` Modell um die Parameter μ und σ einer Gammaverteilung der Meldeverzögerung in Abhängigkeit von der Kalenderwoche der Meldung zu schätzen. Die anderen Parameter wie der Wochentag der Meldung, das Alter, oder das Bundesland des Wohnorts der Fälle haben keinen erkennbaren Einfluss auf die Imputation gezeigt (genaue Ergebnisse hierzu werden an dieser Stelle nicht präsentiert). Wir nehmen also an, dass t_{diff} Gamma verteilt und somit strikt positiv ist:

$$t_{diff} \sim \text{Gamma}(\mu, \sigma),\ \mu, \sigma > 0.$$

In dieser Parametrisierung ist die Wahrscheinlichkeitsdichte der Gammaverteilung gegeben durch

$$f(y; \mu, \sigma) = \frac{y^{\frac{1}{\sigma^2}-1} \exp\left(-\frac{y}{\sigma^2 \mu}\right)}{(\sigma^2 \mu)^{\frac{1}{\sigma^2}} \Gamma\left(\frac{1}{\sigma^2}\right)}.$$

Wir definieren die gleichen Modellgleichungen für μ und σ in Abhängigkeit von der Kalenderwoche der Meldung, x_{kw}, für das `gamlss` Modell:

$$\eta_k = \beta_{k,0} + \beta_{k,1} f(x_{kw}),\ k \in \{\mu, \sigma\}, \tag{1}$$

wobei f eine Glättungsfunktion ist und wir dafür kubische Splines verwenden. Wir schätzen die Paramter $\beta_{k,i}$ basierend auf den Daten mit bekanntem t_{diff} und verwenden die aus dem Modell erhaltenen und von der Kalenderwoche abhängigen Gammaverteilungen um dann die Meldeverzögerung für Fälle ohne bekanntem Erkrankungs-Datum zu imputieren.

Effektive Reproduktionszahl basierend auf dem Erkrankungs-Datum

Basierend auf der Zeitreihe des imputierten und bekannten Erkrankungs-Datums schätzen wir die effektive Reproduktionszahl wie in RICHTER et al. (2020c) beschrieben.

Das von uns benutzte serielle Intervall basiert auf einer Gammaverteilung mit Mittelwert 4.46 und Standardabwichung 2.63. Diese Parameter des seriellen Intervalls ergeben sich aus einer Analyse von 312 österreichischen Quellenfall-Folgefall-Paaren (RICHTER et al., 2020b).

Resultate

Beschreibung der Fallpopulationen

Die Analyse basiert auf den zum Zeitpunkt 29.05.2020 14:00 gemeldeten 16581 COVID19 Fällen (Quellenpopulation). In den Kalenderwochen 20 und 21 gab es überdurchschnittlich viele Fälle, die mit ihrer Labormeldung verzögert in das EMS eingepflegt wurden und bereits ein bekanntes Erkrankungsdatum haben, welches in den Kalenderwochen 10–17 liegt. Diese Fälle ($n = 71$) werden von der Schätzung exkludiert um eine Verzerrung der Verteilung von t_{diff} zu verhindern. Diese Fälle gehen aber sehr wohl in die Darstellung der Fälle nach Erkrankungsbeginn ein (d.h. Epicurve). Von den 13361 Fällen, bei denen das Erkrankungs-Datum bekannt ist werden weitere 73 Fälle exkludiert, weil das Erkrankungs-Datum zeitlich nach dem Labormelde-Datum liegt. Zusätzlich werden 1438 Fälle bei denen der Erkrankungsbeginn am selben Tag wie die Labormeldung angegeben ist (unplausibel) sowie 140 Fälle mit extrem langer Differenz (> 21 Tage) zwischen Erkrankungs-Datum und Labormelde-Datum exkludiert.

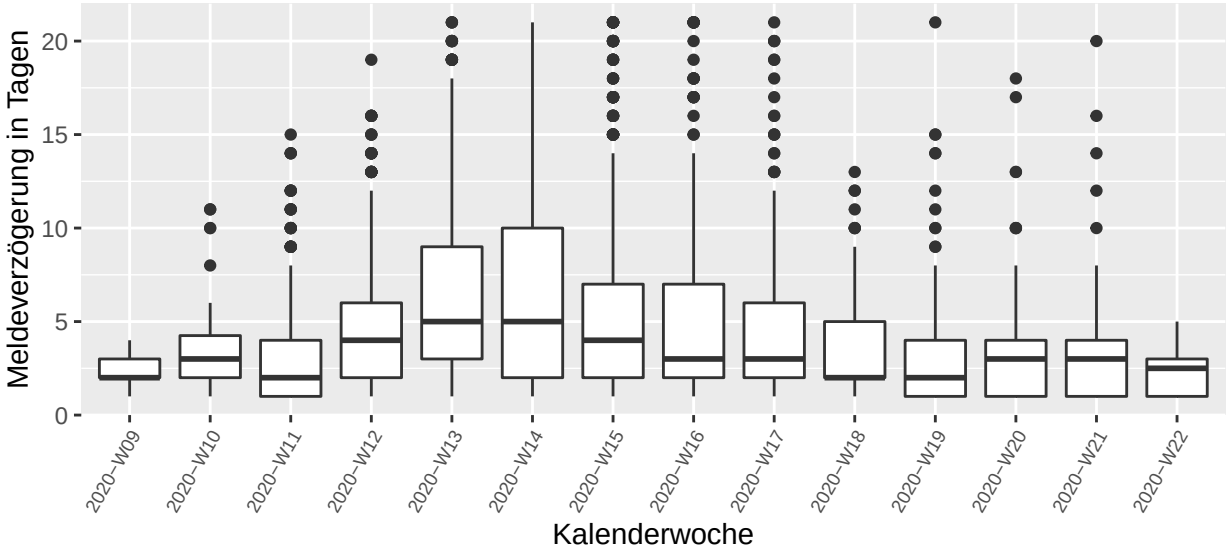
Insgesamt gehen schließlich 11710 Fälle (Studienfallpopulation) in das `gamlss` Modell ein und werden zur Imputation des Erkrankungs-Datums bei 4800 Fällen (ohne Angabe oder mit unplausiblen Angaben zum Erkrankungs-Datum d.h. Fallpopulation mit geschätztem Erkrankungs-Datum) verwendet.

Die Studienfallpopulation weist im Mittel eine Meldeverzögerung von 5.4 Tagen auf. Abbildung 1 und Tabelle 1 stellen die Meldeverzögerung nach Kalenderwoche dar. Wie zu erwarten, zeigt sich in Kalenderwochen mit hoher Fallzahl eine deutlich größere Meldeverzögerung.

Tabelle 1: Anzahl der Fälle gesamt (n), Anzahl der Fälle mit bekannter Meldeverzögerung, Mittlwert und Median der Meldeverzögerung nach Kalenderwoche.

Woche	n	Verzögerung bekannt	Mittelwert	Median
2020-W09	10	8	2.38	2.0
2020-W10	94	88	3.31	3.0
2020-W11	782	608	3.08	2.0
2020-W12	2985	2292	4.47	4.0
2020-W13	4974	3995	6.15	5.0
2020-W14	3172	2367	6.43	5.0
2020-W15	1952	1213	5.01	4.0
2020-W16	759	432	4.91	3.0
2020-W17	471	228	4.79	3.0
2020-W18	316	125	3.39	2.0
2020-W19	258	123	3.56	2.0
2020-W20	337	112	3.51	3.0
2020-W21	271	89	3.49	3.0
2020-W22	129	30	2.47	2.5

Abbildung 1: Boxplotserie der Meldeverzögerung nach Kalenderwoche der Meldung von 11710 Fällen mit bekanntem Erkrankungs- und Melde-Datum.



Imputation des Erkrankungs-Datums

Abbildung 2 zeigt die 16581 Fälle nach Erkrankungs-Datum (bekannt und imputiert). Hierbei handelt es sich um eine einzige Zufallsziehung von t_{diff} basierend auf den modellierten Parametern. Bei jeder neuerlichen Zufallsziehung ergibt sich eine geringfügige Änderung der Form der

Abbildung 2: Epicurve nach Erkrankungs-Datum basierend auf einer zufälligen Stichprobe aus der Verteilung der Meldeverzögerung. Durch die zufällige Imputation bei fehlendem Erkrankungs-Datum können einige Fälle vor dem ersten tatsächlich gemeldeten Fall liegen. Wir weisen ausdrücklich darauf hin, dass dies keine Rückschlüsse auf einen etwaigen "Patient Null" erlaubt und rein statistischer Natur ist.

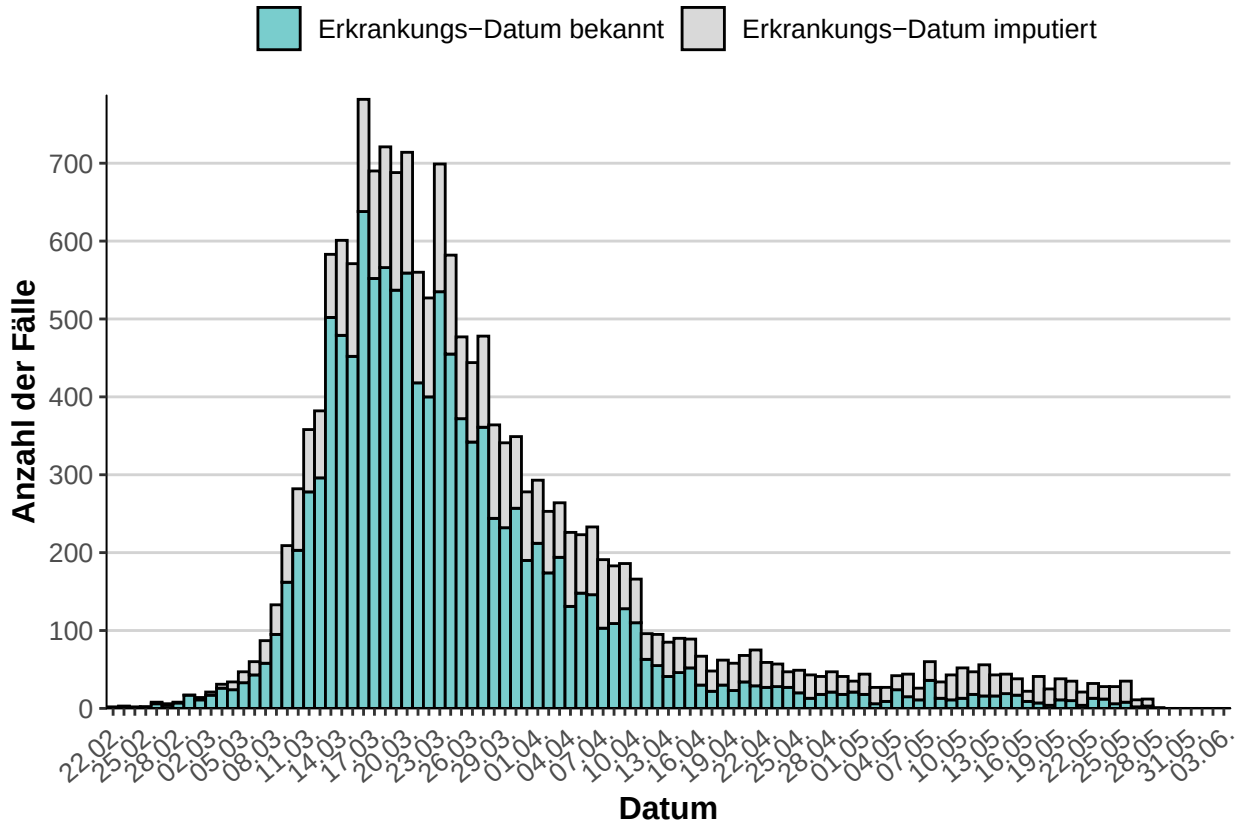
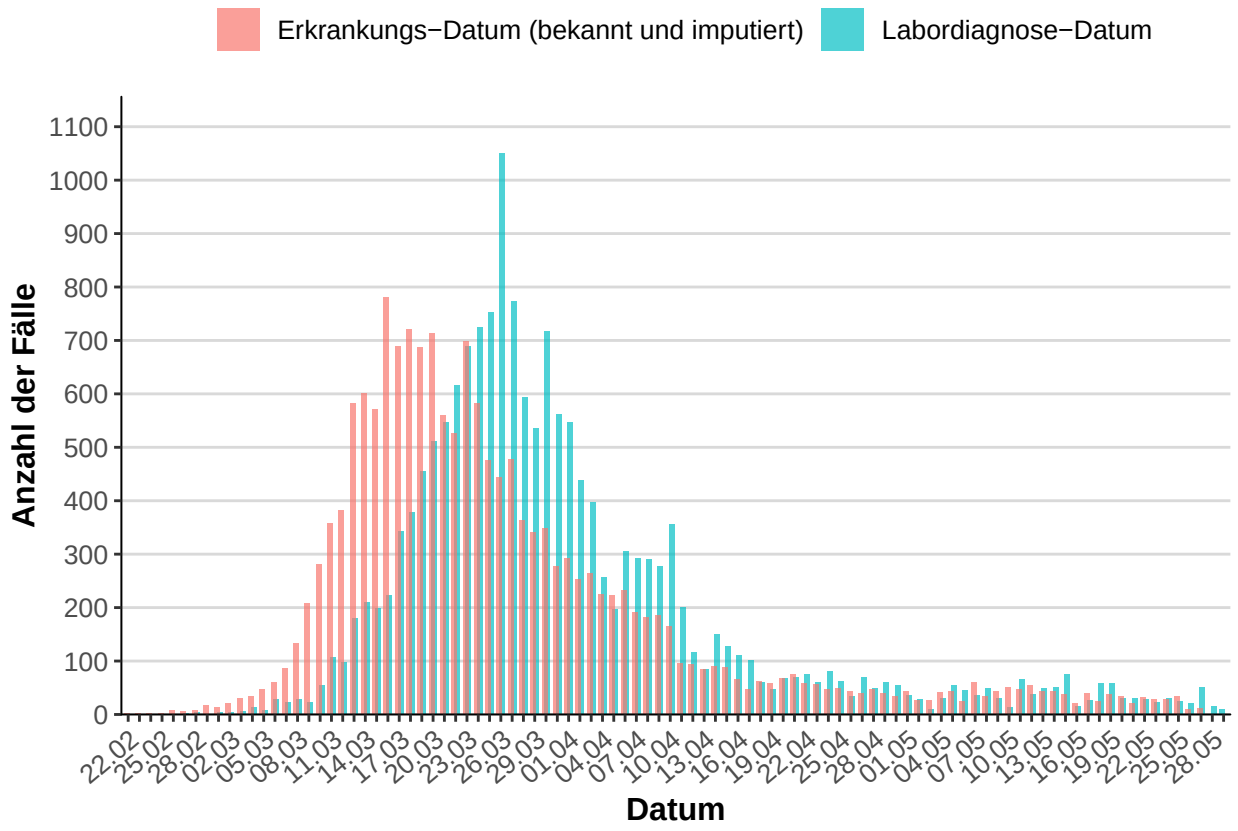


Abbildung 3: Epicurve nach Erkrankungs-Datum (mit bekanntem und imputiertem Datum) versus Epicurve nach Labordiagnose-Datum.

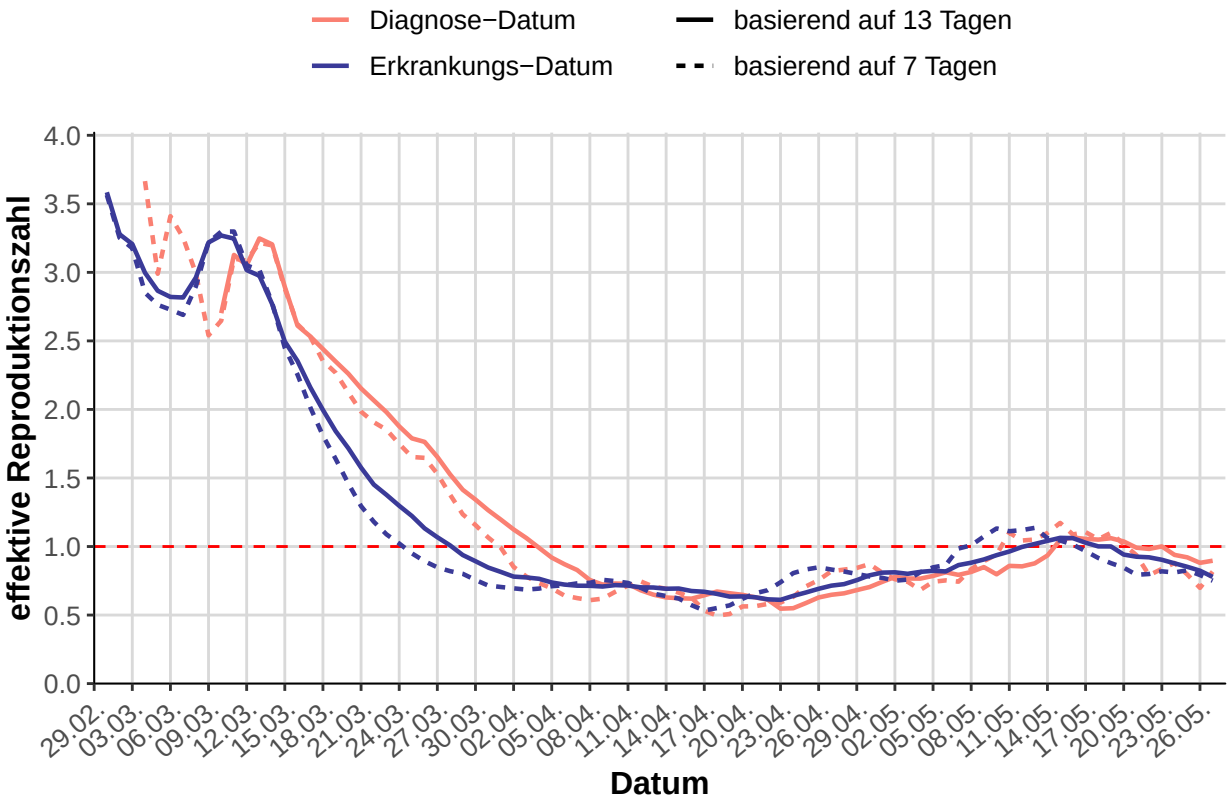


Epicurve. Abbildung 3 zeigt die Epicurve nach Erkrankungs-Datum im Vergleich zu jener nach dem Labordiagnose-Datum. Während die meisten Fälle am 26.03.2020 diagnostiziert wurden, war der Gipfel der Erkrankungen bereits am 16.03.2020 erreicht. Bei 100 Zufallsziehungen war der Tag des Fallzahlmaximum in 99% (99/100) am 16.03.2020.

Effektive Reproduktionszahl basierend auf dem Erkrankungs-Datum

Bisher wurde die effektive Reproduktionszahl mit der Zeitreihe der Fälle nach dem Labordiagnose-Datum geschätzt (siehe zum Beispiel RICHTER et al. (2020a)). Durch die Imputation des fehlenden Erkrankungs-Datums ist es uns nun auch möglich die effektive Reproduktionszahl basierend auf dem Erkrankungsbeginn zu schätzen. Abbildung 4 stellt die effektive Reproduktionszahl unter Verwendung des Erkrankungs-Datums (blau) basierend auf den vorangegangenen 7 bzw. 13 Epidemietagen im Vergleich zur effektiven Reproduktionszahl basierend auf dem Labordiagnose-Datums (rot) dar.

Abbildung 4: Effektive Reproduktionszahl basierend auf dem Erkrankungs-Datum (mit bekanntem und imputiertem Datum) und jeweils 13 bzw. 7 Epidemietagen. Als Vergleich ist auch die effektive Reproduktionszahl basierend auf dem Labordiagnose-Datum dargestellt.



Bemerkung

Derzeit gehen die asymptomatischen Fälle mit einem geschätzten und damit fiktiven Erkrankungs-Datum in die Darstellungen ein.

Referenzen

GÜNTHER, F., BENDER, A., KATZ, K., KÜCHENHOFF, H., HÖHLE, M. (2020): Nowcasting the COVID-19 Pandemic in Bavaria. Preliminary Version 9.

RICHTER, L., SCHMID, D., CHAKERI, A., MARITSCHNIK, S., PFEIFFER, S., STADLOBER, E. (2020a): Epidemiologische Parameter des COVID19 Ausbruchs - Update 17.04.2020, Österreich, 2020. https://www.ages.at/download/0/0/505787c756dc6802696ccd6081cdac50ec214ca5/fileadmin/AGES2015/Wissen-Aktuell/COVID19/Update_Epidemiologische_Parameter_des_COVID19_Ausbruchs_2020-04-17.pdf; letzter Zugriff: 24.04.2020.

RICHTER, L., SCHMID, D., CHAKERI, A., MARITSCHNIK, S., PFEIFFER, S., STADLOBER, E. (2020b): Schätzung des seriellen Intervalles von COVID19, Österreich. https://www.ages.at/download/0/0/068cb5fb9f2256d267e1a3dc8d464623760fcc30/fileadmin/AGES2015/Wissen-Aktuell/COVID19/Sch%C3%A4tzung_des_seriellen_Intervalles_von_COVID19_2020-04-08.pdf; letzter Zugriff: 09.04.2020.

RICHTER, L., SCHMID, D., STADLOBER, E. (2020c): Methodenbeschreibung für die Schätzung von epidemiologischen Parametern des COVID19 Ausbruchs, Österreich. https://www.ages.at/download/0/0/e03842347d92e5922e76993df9ac8e9b28635caa/fileadmin/AGES2015/Wissen-Aktuell/COVID19/Methoden_zur_Sch%C3%A4tzung_der_epi_Parameter.pdf; letzter Zugriff: 21.04.2020.

RIGBY, R.A., STASINOPOULOS, D.M. (2005): Generalized additive models for location, scale and shape,(with discussion). Applied Statistics **54**, 507–554.

Kontakt

Priv. Doz. Dr. med Daniela Schmid
AGES - Österreichische Agentur für Gesundheit und Ernährungssicherheit GmbH
Währingerstraße 25a, 1090 Wien
eMail: daniela.schmid@ages.at
Tel.: +43 (0)5 0555-37304
www.ages.at