Erklärung zur "Charité-Studie"

Jörg Stoye, Professor of Economics, Cornell University

Worum geht es in diesem Dokument?

Ich lege einigermaßen kurz und bündig meine Sicht der Ereignisse dar, welche letztlich zu einem Artikel in der *Bild*-Zeitung, einem Interview mit *Spiegel Online*, diversen weiteren Medienberichten (aktuell >200 mit Gesamtreichweite >700 Mio) und einer Neufassung der Charité-Studie zur Virenlast bei Kindern führten.

Worum geht es in der Charité-Studie?

Die Virenlast in Kindern und Erwachsenen wurde anhand von bereits vorliegenden Daten (sogenannte Gelegenheitsstichprobe) verglichen. Sie ist ähnlich. Das war es eigentlich auch schon. Aufgrund von Beschränkungen, die die Autoren immer offen zugaben (keine repräsentative Stichprobe, Kinder unterrepräsentiert und darum u.U. differenziell selektiert), ist m.E. nicht viel mehr aus den Daten herauszuholen.

Warum dann der Aufstand?

In der Erstfassung des Aufsatzes wurde die statistische Analyse als Test der Hypothese motiviert, dass Erwachsene und Kinder genau dieselbe Virenlast haben. Dies war sicher unglücklich. Zum einen ist diese Hypothese nicht handlungsleitend: Wenn es kleine Unterschiede gibt, wird ein guter statistischer Test tendenziell anspringen und die Hypothese verwerfen, obwohl es eigentlich nicht interessiert. Und genau das ist, zum anderen, passiert. Dies wurde den Autoren genüsslich von diversen Kritikern, z.B. Leonhard Held, David Spiegelhalter und mir selber, vorgerechnet. Die Hypothese wurde im Originalaufsatz nicht verworfen, weil vereinfacht gesagt ein schlechter Test verwendet wurde. Eine schöne Verbildlichung findet sich im Tweet von Christoph Rothe: „Das ist in etwa so, also würde man sich mit einer Lupe auf die Suche nach Bakterien machen, obwohl man ein Mikroskop zur Verfügung gehabt hätte. Wenn man mit der Lupe dann nichts "signifikantes" findet, heißt das erstmal nicht viel."

Zu guter Letzt wurde das Ergebnis im dieser Stellungnahme angehängten Tweet verkündet. In diesem Tweet lesen Sie die Worte „no significant difference". Gleichzeitig sehen Sie eine Punktwolke, in der das statistisch geübte Auge sehr wohl Signifikanz sieht. (Nur zur Sicherheit: Es geht nicht darum, dass die linken Punkte weniger sind, sondern dass sie tendenziell niedriger sind.) Das nahmen Statistiker natürlich auch als sportliche Herausforderung. So kam es zu diversen öffentlichen Kritiken, von denen Bild die meisten (nicht alle!) gefunden hat.

Können Sie die statistische Kritik noch genauer fassen?

Statt einfach eine Regressionsanalyse, z.B. Medianregression, von Virenlast auf Alter durchzuführen, wurden zehn (und in einer zweiten Analyse sechs) Alterskategorien gebildet, nämlich 1-10, 11-20, usw. bis 91-100. Es wurden dann nicht nur Kinder mit Erwachsenen verglichen, sondern alle 45 theoretisch denkbaren Paare gleichzeitig. Falls man aber viele Tests gleichzeitig durchführt, muss man zur Vermeidung falscher „Entdeckungen" die Definition einer Entdeckung im Sinne von statistischer Signifikanz verschärfen. Unter dieser verschärften Definition war dann fast nichts mehr signifikant, woraus in der Conclusio (übrigens nicht ganz zutreffend, wie vor allem Dominik Liebl hervorhob) „gar nichts" wurde. Diese unangemessene und zudem noch unsauber zusammengefasste Teststrategie war nicht das einzige, aber das bei weitem herausragende Problem.

Um zu erklären, warum viele Statistiker mit pointierten Formulierungen reagierten, sollte man bedenken: Zwar war in den Daten letztlich nicht viel zu sehen und der wissenschaftliche Schaden daher gering. Aber grundsätzlich kann man mit der beschriebenen Strategie beliebige Effekte wegrechnen. Auch in einer Regression von Lungenkrebsinzidenz auf Nikotinkonsum findet sich kein signifikanter Effekt mehr, wenn man Nikotinkonsum in hinreichend viele Kategorien aufspaltet und dann alle gleichzeitig paarweise testet.

Was ist dann passiert? (Abgesehen von der Bild-Intervention.)

Wie im wissenschaftlichen Procedere völlig üblich, hat das Charité-Labor die Studie überarbeitet. Hierbei wurden die kritisierten Testmethoden komplett verworfen. Die neue Analyse kombiniert zielgenaue Tests mit einer bayesianischen Regressionsanalyse. Aus meiner Sicht sind die Punktwolken in Figure 6 am eindrücklichsten: Es gibt einen Effekt (die Regressionslinien sind nicht horizontal), aber er ist nicht sehr interessant.

Wer hatte jetzt recht?

Aus rein statistischer Sicht wurde den Kritikern recht gegeben. Das wird auf der allerletzten Seite der Neufassung mit begrüßenswerter Deutlichkeit gesagt. Gerade im Vergleich mit patzigen Rückzugsgefechten, die ich als Fachgutachter schon erlebt habe, möchte ich die klare Stellungnahme loben. Die statistische Analyse wurde auch von Grund auf neu aufgezogen.

Was die Kernaussage der Studie angeht, ändert sich eigentlich nichts. Zwar wurden politische Formulierungen, etwa zu Kita- und Schulöffnung, abgeschwächt. Diese redaktionelle Überarbeitung steht aber in keinem für mich erkennbaren Zusammenhang zur Veränderung der statistischen Analyse.

Insofern war alles ein Sturm im Wasserglas. Dies wussten alle Beteiligten die ganze Zeit. Auch pointierte Formulierungen in meinem Aufsatz sollten hiervon nicht ablenken. Mir war bewusst, dass die Charité-Studie politische Beachtung fand; hätte ich sie für ernsthaft irreführend befunden, hätte ich mich initiativ an die Medien gewendet.

Warum die Kritik öffentlich (auf arxiv.org) zugänglich machen?

Dies ist grundsätzlich ein normaler Vorgang. Auch andere Kritiker machten ihre Einwände fachöffentlich geltend. Hätte ich geahnt, welche Wellen dies schlägt, hätte ich es zwar einerseits vielleicht bei einer email belassen. Andererseits ist der offene Schlagabtausch in der Wissenschaft aber für alle gewinnbringend, und ich will mir eigentlich nicht von Boulevardmedien diesbezüglich den Mund verbieten lassen.

Kann die Öffentlichkeit hieraus lernen?

Ja, darüber wie Wissenschaft funktioniert. Es wird Tacheles geredet: Alle Beteiligten haben mit Sicherheit schon härter ausgeteilt und eingesteckt als in dieser Episode. Zudem sind preprints halt vorläufig und in normalen Zeiten frühestens nach der anonymen Fachbegutachtung nachrichtenwertig. Dass man aktuell nicht so lange warten kann, ist klar. Dieser Konflikt muss m.E. ausgehalten werden.

Worum ging es mir eigentlich?

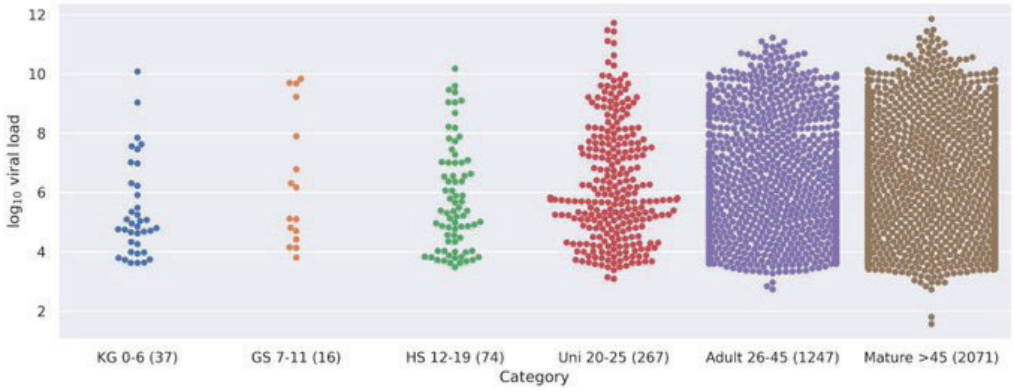
Mal von Besserwisserei abesehen: Ich wollte einen Denkanstoß Richtung „rapid peer review" geben. In der VWL wird dies übrigens z.Zt. praktiziert. Das europäische Center for Economic Policy *Research* veröffentlicht Aufsätze nach einem Schnellcheck, der im vorliegenden Fall vermutlich ausgereicht hätte (die Probleme sind für Statistiker unmittelbar erkennbar), und die Federal Reserve Bank of Minneapolis veröffentlicht z.Z. Aufsätze mit COVID-19 Bezug in ihrer *Quarterly Review* nach vollständigem, wenn auch beschleunigten, Referee-Verfahren.

https://cepr.org/content/covid-economics-vetted-and-real-time-papers-0

https://www.minneapolisfed.org/economic-research/quarterly-reviews

Christian Drosten
@c_drosten

Viral loads by PCR as seen in our laboratories. No significant difference between children and adults. Age categories: Kindergarten (KG), Grade school (GS), Highschool (HS), etc. with age ranges and (counts). bit.ly/SARS-2-load



11.6K 1:54 PM - Apr 29, 2020

6,084 people are talking about this