Grassroots Modeling during the Covid-19 Pandemic

Cecilia Nardini and Fridolin Gross***

* *European School of Molecular Medicine (SEMM) and University of Milan*

** *ImmunoConcept, University of Bordeaux*

Abstract

One of the many peculiar phenomena that the Covid-19 pandemic has brought about is the engagement of non-scientists with specific questions surrounding the interpretation of epidemiological data and models. Many of them have even begun to get involved in the collection, analysis, and presentation of the data themselves. A reason for this might be that the insights that science can provide in a situation of crisis are often inconclusive or preliminary, motivating many people to look for the answers to pressing questions themselves. Moreover, public engagement is facilitated by the easy availability of up-to-date information, of the computational methods to process and analyze it, and of the infrastructure to share and communicate it with like-minded people. This raises epistemological questions about the status of such activities. Can they be considered scientific, and do they meet the standards of scientific inquiry? Or are they harmful because they add to the already loud chorus of voices spreading misinformation and increasing skepticism about the conventional scientific process? We propose to approach this question by looking at a concrete example: A community of active non-professionals has formed in Italy on the software development platform GitHub, where the Italian government's epidemiological data are made publicly available. This represents a well-defined and coherent case study on which detailed information is readily available.

Keywords: Citizen science, Data science, Covid-19, Pseudoscience.

1. Introduction

Nembro, in the province of Bergamo, is the municipality most affected by Covid-19 in relation to the population. We do not know exactly how many people have been infected, but we know that the number of deaths officially attributed to Covid-19 is 31. We are two physicists: one who became an entrepreneur in the health sector, the other a mayor, in close contact with a very cohesive territory, where we know each other very well. We noticed that something in these official numbers did not come back right, and we decided—together—to check.[1]

[1] Cancelli and Foresti 2020 (Claudio Cancelli, Mayor of Nembro, and Luca Foresti, founder of Centro Medico Santagostino).

Argumenta 7,1 (2021): 161-178
ISSN 2465-2334

DOI 10.14275/2465-2334/202113.nar

The Covid-19 pandemic has changed the lives of people all over the world and has altered the way people work, meet, spend their free time, and get healed. And it has also, at least temporarily, changed scientific practice. It is undeniable that research has witnessed an enormous drive to produce results that could help understand the mechanism of transmission of the SARS-Cov2 virus, contain its spread and develop an effective vaccine, all much faster than would have been the case under normal circumstances. At the same time, the traditional scientific method has come under pressure: the conventional peer-review process has been struggling to keep up with the need for fast advancement; the rush to publish often leads to partial results or premature conclusions; and scientific claims are exploited in uncontrollable ways by politicians, the media or other individuals or institutions with an agenda.

In addition to these developments, there has been an unprecedented interest of the public in the details of scientific investigation. Given the direct relevance to their daily lives, people want to understand the numbers that are presented to them by the governments and the media, and to form an opinion on the way in which the pandemic is handled by the responsible institutions. But beyond the interest in existing information and analysis, we observe a significant interest from non-experts to participate in the process of data processing and analysis themselves. This kind of participation is facilitated by the fact that raw data is often publicly available and that it is now easy to obtain state-of-the-art computer tools for analysis and to share and discuss data and results online.

This raises the question whether, and to what extent, these kinds of activities can be considered 'scientific'. More specifically, one may ask to what extent these emerging structures resemble the organization and practices of professional science and whether they have the potential to lead to scientifically respectable outcomes. There is of course a risk that data analysis and modeling carried out outside the realm of conventional science may be used to spread misinformation or to contribute to the acceptance of harmful conspiracy theories. On the other hand, it seems that in the rapidly evolving situation of a global pandemic, conventional science cannot always provide interpretations and predictions quickly enough to meet the needs of the public. Instead of representing an alternative to conventional science, the efforts of non-professional modelers and data analysts may thus be understood as supporting and complementing science in relevant ways. It seems rather obvious to consider these activities as a form of 'citizen science', but at the same time there are clear differences to the paradigmatic examples of citizen science that have been discussed in the literature.

Instead of aiming at an exhaustive overview of the activities of non-professionals related to the pandemic, we decided to focus on a well-defined case study: the community of users of Covid-related data published by the Italian government on the software sharing platform GitHub. The structure of this platform has allowed us to easily follow discussions between members of the community, to track their modeling efforts and analyses, and to identify the outlets that they use to communicate their results to a wider audience. Moreover, the Italian context seems to be particularly interesting, as the situation there was very serious at the beginning of the pandemic, suggesting that the efforts of non-experts are not only driven by curiosity, but by a direct urge to contribute to and accelerate the management of the pandemic crisis.

The paper is structured as follows. In Section 2 we discuss the existing literature on citizen science and place it in the current context of the pandemic to

provide the conceptual basis for framing our case study. We present this case study in detail in Section 3 and discuss it in Section 4. Section 5 offers concluding remarks.

2. Citizen Science

The phenomenon we wish to investigate consists in the increased participation of non-specialists in activities that bear similarities to scientific endeavors. Therefore, it is plausible to consider it within the framework of citizen science. In this section we discuss the way in which the concept of citizen science has been understood in the literature, and we motivate why the idea of citizen science gains particular relevance in the context of the current pandemic crisis.

The term 'citizen science' is relatively new, but it is often pointed out that before the professionalization of science in the 19th century, basically all science was citizen science (Cavalier and Kennedy 2016). The increased attention towards the end of the 20th century and the introduction of the label can thus partly be understood as a reaction to an increasing distance of science from the concerns of the public. The British sociologist Alan Irwin conceived of citizen science as a way of turning science into a more democratic endeavor (Irwin 1995). At roughly the same time, however, the term was also coined in a less politically charged way by Richard Bonney to describe the contribution of scientific data by nonscientists in the context of ornithology projects at Cornell University (Bonney 1996). In line with this, Cooper and Lewenstein (2016) distinguish between two meanings of citizen science: *democratized citizen science* and *contributory citizen science*. An example of democratized citizen science is the involvement of AIDS activists in scientific discussions in the mid-1980s to loosen restrictions on clinical trials and make newly developed treatments available to a wider audience. Prime examples of contributory citizen science are the activities of bird watchers or hobby astronomers who provide the results of their observations to scientific databases.

Contributory citizen science is typically more tightly integrated with science in the traditional sense: hobbyists and laypeople participate in data collection and other data intensive activities that are in turn built on by professional scientists to address relevant problems in their respective fields. However, the contribution of amateur scientists, valuable as it may be, is almost never original, creative, or critically aware. Democratized citizen science, by contrast, is situated at the interface between science and the public and may be seen as a form of interest group advocacy rather than as an epistemic endeavor, although there are cases where people substantiate their concerns by engaging in epistemic activities. An example is the Flint Water Study that involved citizens taking water samples to determine the lead concentration under the direction of professional scientists (Cooper and Lewenstein 2016). The contribution of this type of citizen scientist can be significant in advancing a particular line of research or raising real methodological questions. However, this also makes democratized citizen science more difficult to accept and evaluate this kind of citizen science as a genuine scientific activity.

The polysemous nature of the term makes it hard to find a unifying definition of 'citizen science'. A common theme, however, is that citizen science refers to activities that are carried out in direct interaction with professional scientists. Thus, democratized citizen science aims at convincing scientists of the importance of a particular cause and thereby to exert influence on the direction of research and on scientific policy making. Contributory citizen science, on the

other hand, takes place in the context of projects that are created and supervised by professional scientists. In line with this, the Oxford English Dictionary defines 'citizen science' as "scientific work undertaken by members of the public, often in collaboration with or under the direction of professional scientists and scientific institutions".[2]

Citizen science has mostly been discussed from a sociological perspective, and it has not yet received much attention by philosophers. Existing philosophical discussions have mainly focused on the question whether the contributions of citizen science meet the standards of serious scientific inquiry. For example, Elliott and Rosenberg (2019) discuss three concerns about the quality of citizen science: that citizen science is not hypothesis-driven; that the collected data are of insufficient quality; and that citizen science is biased to the extent that it is politically motivated. They argue that none of these concerns threaten the potential value of citizen science. They point out, for example, that philosophers of science more generally have challenged the notion that all scientific activity must be guided by hypotheses. Thus, scientists themselves switch back and forth between different modes of research, some of which are purely exploratory or data driven. Overall, the activities of citizen scientists are presented as potentially valuable contributions to established science, either by directly adding to scientific projects, by directing scientists to issues of public concern, or by critiquing and modifying established scientific methods.

The Covid-19 pandemic raises pressing questions about the way in which science should be organized in a time of crisis. Answering such questions is not only important for dealing with the current situation, but may also be important in the longer term, as we can assume that similar global crises will occur even more frequently in the future. Philosophers have already given considerable thought to these issues,[3] and some of the most critical problems that have been raised can be understood as pointing to the importance and potential positive impact of citizen science, but also to the risks that such activities may entail in the current context.

First, there is the problem of urgency: science needs to react in a timely manner, and it needs to allocate its scarce resources in the best possible way to produce relevant and reliable results (Reydon 2020). Citizen science projects can help alleviate this problem by contributing to data collection or routine tasks that can be easily outsourced. There have been several examples illustrating the potential of such projects. The Eterna OpenVaccine project enables video game players to "design an mRNA encoding a potential vaccine against the novel coronavirus" (Do Soon and the Eterna Developer Team 2020). Another example is a project launched by UCSF via a smartphone app (Norris 2020), a remote public health study that collects data from participants on their habits and health status to gain insights into the spread of the virus. Lastly, the Rosetta@home (Peckham 2020) crowd-sourcing initiative harnesses the computational power of participants' home computers to find candidates for antiviral drugs. In this latter case the layperson will have the satisfaction of knowing they are contributing to scientific research, even without any original input on their part. Similar examples have

[2] https://www.oed.com/view/Entry/33513?rskey=skqsuT&result=1#eid316619123 (last accessed 07/11/2021).

[3] See the collection of short papers in a recent issue of HPLS, introduced by Boniolo and Onaga (2021).

been discussed in the literature, raising the question whether they constitute genuine cases of citizen science at all (Del Savio et al. 2016).

The second problem is the risk of science developing a "myopic, epidemiology-centric description of reality" (Lohse and Bschir 2020). In other words, there is concern that certain scientific disciplines, such as epidemiology or virology, are being given too much weight at the expense of other relevant fields and perspectives. In response to this, many philosophers have emphasized the need for a more pluralistic approach to the processes of knowledge generation and policy-making that should involve as many stakeholders as possible (Mazzocchi 2021; Ongaro 2021; Leonelli 2021). Clearly, forms of democratized citizen science are one way to address this problem of lack of pluralism, for example by focusing scientists' attention on relevant local contexts or particularly affected population groups. For example, one activist group has written an open letter urging the NIH to include patients with HIV/AIDS in trials of the new SARS/Cov2 vaccines.[4]

Finally, there is the problem of uncertainty and misinformation. Faced with incomplete knowledge and uncertain evidence, scientists have openly disagreed about the best ways to deal with the pandemic,[5] and the accelerated scientific process has led to misuses of results and to the retraction of findings. As a result, large parts of the public have lost trust in the scientific process, which in turn plays in the hands of denialists who question the seriousness of the problem and the need for action to combat the virus (Antiochou 2021; Monasterio Astobiza 2021). Differently from the others, this issue makes the potential role and value of citizen science seem rather ambivalent. On the one hand, citizen science initiatives might have a beneficial impact by critically assessing the way in which science is done, thereby achieving increased transparency and public understanding. On the other hand, there are obvious risks that these activities may lead to the spread of misinformation, adding yet another voice to the already loud chorus that undermines the credibility of conventional science.

In what follows we would like to illuminate these problems and the potential role of citizen science using a concrete case study. This case study shares important similarities with the two types of citizen science identified at the beginning, as it clearly involves a topic of public interest while also mobilizing the skills of many amateur data scientists. At the same time, it seems very different because it looks like a largely self-organized 'grassroots' effort by non-scientists that is not directly linked to the Covid-related projects of the scientific community. Given these features, we think that our case study can contribute to a better understanding of both the value and the potential risks of public participation in the process of knowledge generation and interpretation of scientific evidence. More specifically, we would like to understand whether such activities, when conducted largely independently from established science, necessarily fall into the camp of 'pseudoscience,' whether they lack the quality that other citizen science projects have because of support from professional scientists, or whether they can be understood more positively as indicative of an alternative, more open and inclusive model of scientific research.

[4] https://www.treatmentactiongroup.org/wp-content/uploads/2020/08/covid_19_1273_collins_nih_7_27_20.pdf (last accessed 07/11/2021).

[5] For an example consider the exchange between Ioannidis and Lipsitch (Ioannidis 2020; Lipsitch 2020).

3. The Case Study

On 21 February 2020, ten small municipalities in Lombardy were quarantined after the discovery of a local hotspot of the new Coronavirus disease. In the following weeks, the pandemic took the country by storm. Italy was the first western country to feel the effects of the new virus and the first to enter a nationwide lockdown on 9 March. The Italian Civil Protection Department (DPC) was engaged since the beginning of the emergency and started releasing daily communications of the epidemiological situation by assembling data from all the different regional health systems. These reports were made public every day at 6PM in a press conference and subsequently published in a .pdf document on the official website of the DPC. On 4 March 2020, the open data association onData[6] began automatically scraping these files to publish the same data on the software development platform GitHub in a format more suitable for further computational analysis[7], while petitioning the Italian government to publish the data in an open, machine-readable format directly from the DPC. On 7 March 2020, the DPC opened its own GitHub repository, where the data contained in the daily report were published daily after the press conference in machine-readable format and under a Creative Commons license. Since 25 June 2020, the data has been released directly by the Ministry of Health, but the open data repository continues to be curated by the DPC.

A community of users quickly formed around both the first unofficial repository and the official DPC repository, using the data for a variety of purposes, such as personal or publicly available spreadsheets or dashboards to monitor the progression of the pandemic. The GitHub 'Issues' system provides a convenient way to map and analyze this community. 'Issues' are online discussions usually related to queries, clarification requests or bug reports that can be opened on a repository by registered users. Although they are mainly intended as a means of communication between the maintainers and the users of the repository, issues can—and often do—become a place of communication and exchange among a broader community of users, especially when the repository is public. This has happened to a great extent with the 'Issues' section of the DPC repository: with several issues opened every week and only two official maintainers actively taking care of the DPC repository—users @umbros (Umberto Rosini) and @pierluigicara (Pierluigi Cara)—queries and clarifications from external users are often answered by other external users, leading in some cases to long and intense discussions.

3.1 Methods

To analyze the community of users of the GitHub repository of the Italian Government, we surveyed the approximately one thousand issues published there from the date of its creation (7 March 2020) to April 2021 and selected those that met our criteria for interest/significance.

Issues generally fall into two broad categories. A subset of issues are opened to signal minor inconsistencies or errors in the data that the users can quickly spot because of the type of automated analysis they perform on the data. Issues of this

[6] https://ondata.it/ (last accessed 06/11/2021).
[7] https://github.com/ondata/covid19italia (last accessed 06/11/2021).

type usually receive a reply from the maintainers of the repository and are sometimes passed on to the DPC/Health Ministry to prompt a correction.

We chose to focus on a second, more philosophically interesting kind of issues that contain open-ended, methodological discussions. Here, maintainers intervene sparingly, if at all, while the participation of other members of the community is often lively. Furthermore, the questions posed in these issues often remain unresolved, which provides an interesting parallel to open research questions in traditional science.

After identifying the most interesting issues in this way, we tracked the users participating in the discussions to identify possible modeling and analysis efforts beyond their contribution to the repository. This was easy when the users published their work in their own GitHub repository or on a personal website linked to their GitHub profile. In some cases, however, it was impossible to track down this additional work from a user's GitHub profile, even if they had mentioned it in the discussion.

In the following, we will first present the individual issues that we believe are relevant for the context of our paper. Afterwards, we will present four examples of larger projects and several examples of individual users who engaged in modeling based on the data provided by the repository.

3.2 Findings

3.2.1 Individual Issues

Our first step in analyzing the community of GitHub users consisted in a survey of individual issues in the repository.

Of all the topics we captured for follow-up and used to identify the users' modeling efforts, we describe below a selection of issues that we believe can provide a representative overview of the type of discussions taking place among community members.

- Issue #577 concerns the data collected in the field 'tamponi' (swab tests) of the published data. The discussion clarifies that in different Italian regions data are generated differently. For example, in some regions they include all swabs, while in others only swabs that have already been analyzed are counted.
- Issue #587 concerns the estimation of R_0 and R_t, the initial and the current reproduction number of the virus. This is a very interesting issue because several of the participants propose their own analysis of these metrics. For example, users @alessandroNa, @Riccardocominotti, @LucaZeta, @brunocaniglia and @mpreitano present their methods for a simplified estimation of R_0, and they suggest possible improvements to each other. Additionally, user @Pivone presents his detailed analysis based on several indicators constructed from the DPC data (this user will be treated more in depth in Section 3.2.3).
- Issue #821 concerns two new fields that were added to the dataset (and later removed): 'Casi da sospetto diagnostico' and 'Casi da screening' (cases found due to diagnostic suspicion or via screening, respectively). Participants in the discussion debate the correct interpretation of the two new fields and provide evidence that the definitions of the two metrics are interpreted differently by different regions, leading to inconsistencies in the data. The distinction is considered relevant because of the different probabilities of finding an active (contagious) case by the two different methods. In the

discussion, two different interpretations are proposed: according to users @Paulsword and @Rabelaiss the first category (diagnostic suspicion) includes also cases found via contact tracing, so that cases in this category are more likely to be active spreaders of the infection. According to user @vienne, by contrast, contact tracing falls in the second category (screening), which is therefore the category that has the greater likelihood of including infection hotspots. Both sides of the discussion support their point of view by citing the national or regional health authorities and data. Ultimately, however, the issue remains unresolved, as further clarification from the Ministry of Health is still pending at the time of writing, and the new fields are removed from the dataset anyway in a later revision.

- Issue #864 features a very interesting and long discussion on the definition of some of the main quantities provided as open data on the repository, the fields 'Casi testati' (people tested) and 'Tamponi' (samples tested). These definitions are crucial because the two measures are used by health authorities and the media as a basis to calculate the daily incidence figures. In the discussion it is noted how ambiguity or inconsistency in definition may lead to systematic overestimation or underestimation of the daily incidence. The arguments put forward by the participants are valid, but they leave open the question of whether the competent health authorities have made the same considerations.
- Issue#892 highlights an apparent inconsistency in the trend of the number of recovered patients vs new cases. The claim is backed by a graphical analysis of the data in question. However, there is no official acknowledgment of the anomaly.
- Issue #977 is a very debated one in which at least two interesting questions are analyzed. The starting point is the introduction of a new field in the dataset ('Ingressi del giorno in terapia intensiva', daily new entries in ICU), and participants discuss the relation between this new field and other quantities in the dataset. A second question that appears in the same issue is the usefulness of an index, introduced by user @CT-igiul, based on the relative variation of current positive cases. User @Rabelaiss argues that this index does not give a useful picture of the evolution of the pandemic, while user @Doc73 points out that it bears similarities with the technique of derivative control used in industrial control systems. This issue is interesting in terms of its content, but also because it represents an example of methodological discussion in which the community productively engages with the work of one of its members.
- Issue #1005 concerns the observation of suspicious simultaneous spikes in weekly averages of deaths, cases, and tests. Some explanatory hypotheses are proposed, but again there is no official acknowledgment of the problem.
- Issue #1136 is opened by user @AntonioB1976 as a fact-checking request into a Covid-19 denialist's claims on Facebook that the number of new positive cases reported daily by authorities includes repeated tests of already know positive cases. None of the maintainers intervene to make an official statement, but some of the most active users provide data and observations to refute the denialist's claims.

The issues we have singled out constitute a representative sample of the kind of interaction and dialogue that takes place in this community. As can be seen,

the general tone is altogether different from other social media forums. Issues are usually opened with a precise methodological or data-related question in mind, and the answers are not purely opinion-based, but are usually supported with references to scientific literature, to health authorities or directly with data and analysis results. The debates are rational, and the common goal appears to be that of gaining a better understanding of the underlying issue or of the data, rather than to convince others of one's own opinion. On the other hand, the discussions resemble those on other social forums in that their impact is limited to the participants or, at best, to other interested members of the community. In some instances (e.g., in issue #821 mentioned above) official maintainer @umbros intervened to say that he would submit a query to the Health Ministry for clarification, but in all cases were this was done, the official response, if ever given, was not reported back on the repository.

3.2.2 Larger Projects

After looking at individual issues, we proceeded to track participants through their GitHub profiles and assessed whether there was any research projects available on their GitHub profile or otherwise reachable via links from there. We were able to identify several web dashboards fed with the DPC data from the repository, and we included them in our analysis if there was evidence of original research content beyond mere reporting or visualization of the data. In the following we will detail the main modeling efforts that we identified in this way.

EpiDataItalia

According to their website,[8] EpiDataItalia is a study and research group formed by three self-funded volunteers (a data scientist, a biologist by training and a musician who is an amateur mathematician/statistician). It seeks to offer analysis and forecast on the COVID-19 pandemic with particular attention to the situation in Italy. Apart from directly posting on their website, they have published their results and analyses in the news magazine *L'Espresso* and as preprints on open access repositories, such as *Zenodo*.

One part of their project consists in processing and visualizing the data provided at the national level by the Italian government and at the international level by Johns Hopkins University. However, they go beyond mere reporting of data by pursuing their own scientific questions, for example the correlation between air pollutants and COVID-19 cases in the Lombardy region. They also compared in detail different methods for calculating the effective reproductive number R_t, proposed a new way of estimating the case fatality rate, and investigated the consequences of different vaccination strategies using mathematical models.

ilsegnalatore.info

This is mostly a scientific news website, originally created to provide controlled, verified pieces of news and information on the pandemic. The main author of the site is a physician, Paolo Spada (user @paulsword on GitHub), who is an active participant in many of the GitHub issues that we analyzed. Since March 2020 he has published a detailed daily report with infographics on the DPC data. The report, published both on the website and on his own Facebook page, is followed by thousands of citizens and has attracted national-level attention with an

[8] https://www.epidata.it last accessed 06/11/2021.

interview in the magazine *Panorama* (Bonaccorso 2020) and several interviews on national television.[9]

The detailed daily report is interesting because it contains elaborations that go beyond the mere visualization of the time series of data. For instance, in a graph recently added to the report, trends in incidence rates, rates of ICU admission and fatality are plotted against vaccination coverage in the 60+ age group, with the assumption that the latter two values should be lower than in the previous waves because of the protection offered by vaccination to the most vulnerable age group.

In addition to the daily reports, Spada has published around 20 articles on the website. Some are mainly scientific communication articles: for example, there is an explanation and commentary on a graph published in *JAMA* depicting how the probability of detecting an infection with different tests varies over time,[10] and an explanation of the meaning of the various indicators that are communicated daily by newspapers. However, some go beyond scientific communication by including a critical review of the available data, often in the light of current scientific literature. For instance, in an early article[11] he compared the predictions of the *SIR* model with the actual data to show an apparent overestimation of recovered patients in the data reported by the Lombardy region. Another one, "Oltre l'R_t",[12] ("Beyond R_t"), proposes and evaluates the use of the weekly average of the percentage variation of new cases as a proxy measure for R_t. The interest of this proxy measure is that it is a value that is readily available from the data up to the current day, unlike R_t itself which has a considerable lag because it needs to consider the time interval between the infection and the onset of symptoms and between the onset of symptoms and the diagnosis. The comparison between the two values (weekly average of the percent variation vs. R_t shifted back in time) is shown daily in the reports on the website and there is indeed a strong correspondence between the two curves.

OpenCovidItaly initiative

OpenCovidItaly is one of the data users that we identified starting from the first unofficial OnData repository. It is a blog/study group that published several articles and analyses between May and August 2020.

There is neither a detailed description of the group's structure nor a listing of the individual participants, but the "Perché" (Why) section of the blog[13] explains that the main motivation of the initiative is to provide open data on the pandemic through scraping and collecting data from various sources.

The first posts are indeed just data presentation, providing a breakdown of the data about deaths in some Italian regions. However, subsequent posts include some elaborations on the presented data. In particular, there is a methodological

[9] For instance, https://ilsegnalatore.info/frontiere-raitre/ or https://ilsegnalatore.info/tg5-ore-20-mediaset-4/ (last accessed 06/11/2021).

[10] https://ilsegnalatore.info/una-figura-e-meglio-di-tante-parole/ (last accessed 06/11/2021).

[11] https://ilsegnalatore.info/i-pazienti-dimenticati-nei-conti-della-lombardia/ (last accessed 06/11/2021).

[12] https://ilsegnalatore.info/oltre-allrt/ last accessed (06/11/2021).

[13] https://opencoviditaly.netsons.org/why/ (last accessed 06/11/2021).

article[14] explaining the risk of using data that are not yet consolidated, and an explanation[15] of the epidemiological indicator R_t with an in-depth analysis of how it can be estimated from data that are constantly under accrual.

Currently, they use their twitter profile[16] to publish a weekly forecast of the value of R_t. They then proceed to confront this forecast with the official figure released by the National Health Institute the following day. This is a particularly interesting example of the interaction of grassroots modelers because this forecast for R_t is obtained using a web application published by another participant in the community (user @vi-enne mentioned below), fed with data provided by the OnData collective, the original unofficial source of data which currently is still providing some finer-granularity data that would otherwise be unavailable in machine-readable format.

Vittorio Nicoletta

User @vi-enne (Vittorio Nicoletta) is a computer scientist active on twitter with the handle @vi__enne.[17] He has a public repository[18] in which he answers some frequently asked questions on Covid with explanations, data, and literature references. He also published an analysis dashboard (a public Google Drive worksheet) for forecasting the level of risk and transmission in the different Italian regions. Finally, he created a web application[19] that allows any web user to estimate the value of R_t from the official data or from any dataset they provide in .csv format. The app uses the models available in the EpiEstim R software package and allows the user to set some analysis parameters and choose between the four available models for estimation. This is the application mentioned above that is used by the OpenCovidItaly group to estimate their weekly R_t forecast.

3.2.3 Others

Besides the more prominent examples that we have mentioned so far, we have identified several less systematic but still noteworthy modeling efforts.

- GitHub user @alexamici (Alessandro Amici) has a repository[20] of Python notebooks that are updated daily with data and short-term forecasts at national and regional levels. He also has a blog[21] that he updated between March and October 2020 with statistical and data analytic considerations.
- Users @littleark (Carlo Zapponi) and @leeppolis (Simone Lippoli) are the founders of *Visualize News*, a group of computational designers with an interest in data visualization. They curate an infographics dashboard on Covid[22] which also includes some elements of original analysis, e.g., the section "How is today's situation compared with the first wave?"

[14] https://opencoviditaly.netsons.org/cosa-succede-quando-si-utilizzano-dati-non-consolidati/ (last accessed 06/11/2021).
[15] https://opencoviditaly.netsons.org/erreti-leggermente-maggiore-di-uno/ (last accessed 06/11/2021).
[16] https://twitter.com/OpencovidM (last accessed 06/11/2021).
[17] https://twitter.com/vi__enne (last accessed 06/11/2021).
[18] https://github.com/vi-enne/FAQ_covid19_ITA (last accessed 06/11/2021).
[19] https://vienne.shinyapps.io/rt_estimation/ (last accessed 06/11/2021).
[20] https://github.com/alexamici/covid-19-notebooks (last accessed 06/11/2021).
[21] https://naturalstupidity.ghost.io (last accessed 06/11/2021).
[22] https://coronavirus.visualize.news/ (last accessed 31/05/2021).

- User @fotografAle (who is, according to his profile, a professional photographer) was active in some of the issues analyzed above. In one of them he attached a plot and referenced an analysis that resulted from a deep learning forecast model he created. Unfortunately, the model (which he says is based on a convolutional neural network) is not published in his GitHub profile and does not appear to be publicly accessible, so this mention in one of the issues is the only reference to its existence.
- User @heyteacher has a GitHub repository[23] with a machine learning project for forecasting the evolution of the pandemic. Unfortunately, it appears to be abandoned (last updated in June 2020), and there is no way to assess it now. The dashboard[24] by the same author is still updated but contains only data visualization.
- User @LucaZeta is another very active participants in many of the issues. He has created a dashboard[25] that looks quite confusing. There are lines in the graph labeled as 'analysis' but no indication at how they are derived.
- User @vitop72 has a public repository named Covid19 Italy Report[26] updated between April and June 2020 with weekly reports (in .pdf format) that contain a graphical elaboration of the various pandemic-related indicators and a forecast for the coming week.
- User @CT-igiul (Luigi Tomaselli) was very active in some of the issues. He publishes a blog[27], still updated as of May 2021, with some analyses and elaboration; in particular, he has developed an indicator based on the daily relative variation of current positive cases.
- User @Pivone participated in one of the issues posting details and some results of his analysis.[28] He developed some indicators for the development of the pandemic, such as a simple estimate of R_0 and a linear regression. He also analyzed the ratio between home quarantined patients and patients in hospitals for various regions. This allowed him to conclude that in the first months of the pandemic in Lombardy mostly only people with severe symptoms were tested, a fact that has since been officially recognized.
- User @SilvioCaggia (Silvio Caggia) also shared his analysis in the context of an issue.[29] His model is a Google Drive spreadsheet document[30] with graphs and visualizations of the DPC data, but there are hints of original analysis, for instance the sheet 'Qcomparativo' which, as he writes in the issue, is an attempt to analyze fatality rates in different regions.

All the models that we examined can be placed on an axis where, on one end, there are personal research efforts that users pursue in isolation and are reluctant to share (e.g. the projects of users @fotografAle, @Pivone and

[23] https://github.com/heyteacher/sam-forecast-automation-covid-19-ita (last accessed 06/11/2021).
[24] https://heyteacher.github.io/COVID-19/#/ (last accessed 06/11/2021).
[25] https://covid19.zappi.me/ (last accessed 06/11/2021).
[26] https://github.com/vitop72/Covid19-Italy-Report (last accessed 06/11/2021).
[27] https://www.luigitomaselli.com/ (last accessed 06/11/2021).
[28] https://github.com/pcm-dpc/COVID-19/issues/587#issuecomment-637168807 (last accessed 06/11/2021).
[29] https://github.com/pcm-dpc/COVID-19/issues/759 (last accessed 06/11/2021).
[30] lhttps://docs.google.com/spreadsheets/d/11S6KS8lpYq_rNYdf4uqZhKgmPSmHvG9s7S0O-dD4PH4/edit#gid=1092157180 (last accessed 31/05/2021).

@SilvioCaggia described in Section 3.2.3), while, on the other end, there are public dashboards or blogs whose main motivation is scientifically supported public communication (the cases of Visualize News and ilsegnalatore.info are the most obvious examples).

Between these two extremes, some projects are shared with a less wide audience in mind, that is, with a community of experts and insiders. This is the case, for instance, of the application developed by Vittorio Nicoletta for the estimation of R_t, of many of the elaborations by the OpenCovidItaly initiative, or of the blog curated by user @alexamici. For these kinds of projects, the main distribution channels outside of GitHub are traditional social media, such as Twitter. Indeed, from a brief analysis of this informal 'citation network' we found a certain level of interplay between these projects.

In the next section we will consider what our findings entail for the original questions we set out to examine.

4. Discussion

Our case study provides detailed insight into a community of non-professionals who engage in the presentation and analysis of data related to the current pandemic. How seriously should one take this kind of activity? Can it be called 'scientific', or is it more the hobbyhorse of a group of 'data nerds' who play scientists to entertain themselves while they are stuck at home? While it is difficult for us to directly assess the scientific merit of the analyses and models proposed by the members of the community, we can at least look at their practices and interactions as revealed in the GitHub discussions and compare them to genuine scientific activities.

There are several ways in which what we observe resembles the practices of professional scientists (or at least the normative ideal of science). First, the discussions are constructive and rational and very different in style from the kinds of discussions one can witness on other social media platforms. Participants usually share a common goal of better understanding a particular phenomenon, concept, or methodological issue, e.g., how well quantitative measurements provided by public institutions reflect the true dynamics of the pandemic. And they do not appear to be pursuing their activities for financial or personal gain. Second, the members of the community engage in open sharing of their results, resources, methods, and concepts that they use. This kind of collaboration is facilitated by the fact that data and software code are typically made openly available on GitHub, and we observe that in some cases participants use other participants' results for their own projects and build on them. Taken together, this suggests that members of this community adhere to the norms typically associated with the ideal of scientific conduct (Merton 1942; Anderson et al. 2010). Furthermore, we can find a certain degree of continuity between the activities of the community and established scientific research, in that the members of the community explicitly refer to scientific resources and make use of concepts and statistical tools from epidemiological research (e.g., by using the same software packages that are also used by professional scientists). And it does not appear that their work advocates doctrines that are in tension with established beliefs of the relevant scientific fields. All this suggests that the activities of the community cannot simply be dismissed as pseudoscientific in the same way as, for example, the alternative theories and models of climate change deniers (Hansson 2017).

On the other hand, there are clear differences between what we observe in the GitHub community and established science, for better or for worse. Overall, the activities of the community are less coherent, as everyone works mainly on their own problems and analytical tools, despite intense discussions and occasional sharing of resources. Moreover, there are no agreed standards or measures for peer control. Instead, users present their work to others and allow it to be critiqued on a purely voluntary basis. Finally, there are no restrictions on entry into the discussion. Participation in conventional science is typically restricted to individuals who have an accepted degree of qualification (e.g., a PhD) and are affiliated with an official institution (e.g., a university). The GitHub community, by contrast, is in principle open to anyone who has an account. Such openness carries an obvious risk of lowering the quality of the output of the community. However, this feature can also be seen positively as it increases the diversity of participants and can make research more productive and balanced, which, as we have discussed in Section 2, is especially relevant in the context of a social crisis such as a pandemic.

We think that the two standard modes of citizen science, democratized citizen science and contributory citizen science, do not really capture what we observe in our case study. First, the activities of the GitHub community are completely bottom-up and self-contained, i.e., carried out without any direct involvement of professional scientists. Secondly, the citizen scientists in our case study do not want to influence the scientists, but to take matters into their own hands: the community members would not be content to work within the framework of established methods, as part of their activities is precisely to question and criticize these methods.

Of the citizen scientists we encountered in our study, many are motivated by a desire to improve their personal understanding of the situation by analyzing the data on their own. Some users are skeptical of the way valid scientific results are interpreted by the media and disseminated to the public, and therefore develop their own measures and analyses to understand and critically evaluate the news. The words of user @Pivone, mentioned in Section 3.2.3, exemplify this attitude: "credo che questo set di dati aggregati [...] permetta di fare uno screening ragionato e serio delle notizie e di rigettare le molte cose fuorvianti che sono state dette e scritte in proposito"[31] (I believe that this dataset allows for a reasonable and serious screening of the news and to disprove the many misleading things that have been said on the matter).

In other cases, however, there is real dissatisfaction among modelers because they feel that conventional science and government agencies are overwhelmed and cannot respond with the necessary care or transparency in their communications. An example is the recurring issue of the right way to determine R_t, the reproduction number of the virus. There is no consensus on how best to estimate this number, but it appears to be crucial because it expresses in a simple way where the pandemic is headed. Thus, the members of the community respond to the need to increase transparency around this issue, feeling that the scientific community and government institutions at various levels often send conflicting messages. For example, user @PaulSword developed an alternative measure of the progress of the pandemic in his articles on ilsegnalatore.info, mentioned in

[31] https://github.com/pcm-dpc/COVID-19/issues/587#issuecomment-642287928 (last accessed 06/11/2021).

Section 3.2.2, based on weekly average variation of new cases. This metric, while being easy to compute and based on current data, provides a very good approximation of the official estimate of R_t. The good fit between the two measures is shown on the ilsegnaltore.info blog with weekly updated charts. His question as to why the authorities do not take this simplified model into account seems legitimate, especially since the measure he developed has the advantage of being almost real-time, unlike the official estimate, which can only be calculated with a delay of two weeks.

More generally, the focus on R_t, and on similar metrics that capture the progression of the pandemic offers important insight into the ultimately social motivations behind the research efforts of the community. Fostering collaborative and safe behavior in citizens through clear and accurate scientific communication is a strong motivation behind the effort of some of the modelers we studied. As pointed out earlier, this *bona fide* sentiment seems far removed from denialist positions or attempts to propagate conspiracy theories. Most members of the community appear to have professional experience in dealing with complex data and are therefore aware of the methodological pitfalls that can affect data-based decision-making in situations such as this one, where a large amount of heterogeneous data must be collected quickly, and the data collection pipeline had to be set up hastily and with little oversight.

An example that supports this idea comes from an altogether different case: the COVID Tracking Project in the U.S.,[32] which is a data collection initiative that was launched by the news magazine *The Atlantic* in March 2020 out of dissatisfaction with the data the U.S. Centers for Disease Control and Prevention (CDC) were making publicly available (Meyer and Madrigal 2021). Over a few months the project, which was based on the data collection effort by hundreds of citizen volunteers, became the most complete data source about COVID-19 in the U.S., being used by *The New York Times*, Johns Hopkins University, and two presidential administrations, as well as being cited in more than 1,000 academic papers, including major medical journals like *The New England Journal of Medicine*, *Nature*, and *JAMA*. The project's founders, Meyer and Madrigal (two journalists), emphasize the importance of understanding the way data is collected and organized in order to be interpreted correctly:

> The scientists at the CDC clearly have far more expertise in infectious-disease containment than almost anyone at the COVID Tracking Project or *The Atlantic*. But we did spend a year grappling with the limitations of the system (Meyer and Madrigal 2021).

The example of the COVID Tracking Project is, we believe, a success story that provides a glimpse of what the GitHub community could have looked like if somebody with the necessary power had managed to coordinate the efforts, the data skills, and methodological expertise of the participants.

For it must be acknowledged that while the members of the GitHub community have clearly achieved useful results, their activities remain somewhat fragmented and do not seem to be having the kind of impact that a better organized and streamlined project could have achieved.

[32] https://covidtracking.com/ (last accessed 06/11/2021).

5. Conclusion

In this article we examined the activity around the GitHub repository where the Italian government publishes data related to the Covid19 pandemic. What we have discovered is a peculiar kind of citizen science in which lay people try to improve the kind of information they get from official sources. We have argued that these activities can at least partly be considered scientific, but they are different from other forms of citizen science because they do not rely on the direct involvement of professional scientists. Rather, we observe that the citizen scientists in our case study attempt to circumvent or 'short-circuit' the usual flow of information that reaches the public via science or the media.

Obviously, our analysis provides only a short glimpse into a phenomenon that might itself be transient and dependent on the dynamics of the pandemic. Furthermore, we were not able to systematically track many of the activities of community members that took place outside the GitHub platform. Thus, it may be promising to map the informal 'citation networks' of citizen scientists across social media platforms such as Facebook and Twitter and compare them to the organization of established science.

Despite its epistemic shortcomings, we see the community described in our case study as a positive example that avoids some of the risks typically associated with public participation in controversial scientific topics, while at the same time exhibiting greater openness and diversity, a feature that seems particularly relevant in the current crisis.

We started our article with a quote from Claudio Cancelli, the Mayor of Nembro, and his fellow citizen scientist Luca Foresti. In the same article they sum up the particular requirements of the current situation:

> We are in the midst of an epoch-making event and to fight it we need credible data on the reality of the situation, disclosed transparently among all the experts and people who have to manage the crisis responsibly. Based on these data we can understand and decide what is right to do when it is required (Cancelli and Foresti 2020).

This call should be understood in its broadest sense, as we are all citizens involved in the responsible management of this crisis. Therefore, we believe it reflects the unprecedented momentum that has led many citizen scientists to commit their time and efforts to contribute to a better understanding of the Covid-19 pandemic.

References

Anderson, M.S., Ronning, E.A., Vries, R.D., and Martinson, B.C. 2010, "Extending the Mertonian Norms: Scientists' Subscription to Norms of Research", *The Journal of Higher Education*, 81, 3, 366-93.

Antiochou, K. 2021, "Science Communication: Challenges and Dilemmas in the Age of COVID-19", *History and Philosophy of the Life Sciences*, 43, 3, 87.

Bonaccorso, M. 2020, "Il Medico dei Numeri del Covid-19", *Panorama*, Mar 31, https://www.panorama.it/news/salute/il-medico-dei-numeri-del-covid-19 (last accessed 07/11/21).

Boniolo, G. and L. Onaga 2021, "Seeing Clearly through COVID-19: Current and Future Questions for the History and Philosophy of the Life Sciences", *History and Philosophy of the Life Sciences*, 43, 2, 83.

Bonney, R. 1996, "Citizen Science: A Lab Tradition", *Living Bird*, 15, 4, 7-15.

Cancelli, C. and Foresti, L. 2020, "The Real Death Toll for Covid-19 is at least 4 Times the Official Numbers", *Il Corriere della Sera*, English version, March 26, https://www.corriere.it/politica/20_marzo_26/the-real-death-toll-for-covid-19-is-at-least-4-times-the-official-numbers-b5af0edc-6eeb-11ea-925b-a0c3cdbe1130.shtml (last accessed 30/11/2021).

Cavalier, D. and Kennedy, E.B. (eds.) 2016, *The Rightful Place of Science: Citizen Science*, Tempe: Consortium for Science, Policy & Outcomes.

Cooper, C.B. and Lewenstein, B.W. 2016, "Two Meanings of Citizen Science", in Cavalier and Kennedy, 51-61.

Del Savio, L., Prainsack, B., and Buyx, A. 2016, "Crowdsourcing the Human Gut: Is Crowdsourcing Also 'Citizen Science'?", *Journal of Science Communication*, 15, 3, A03.

Do Soon and the Eterna Developer Team 2020, Eterna OpenVaccine, https://eterna-game.org/ (last accessed 06/11/2021).

Elliott, K.C. and Rosenberg, J. 2019, "Philosophical Foundations for Citizen Science", *Citizen Science: Theory and Practice*, 4, 1, 9.

Hansson, S.O. 2017, "Science Denial as a Form of Pseudoscience", *Studies in History and Philosophy of Science*, Part A 63, 39-47.

Ioannidis, J.P.A. 2020, "The Totality of the Evidence", *Boston Review*, 26, 22-30.

Irwin, A. 1995, *Citizen Science: A Study of People, Expertise and Sustainable Development*, London: Routledge.

Leonelli, S. 2021, "Data Science in Times of Pan(dem)ic", *Harvard Data Science Review*, 3, 1, doi: 10.1162/99608f92.fbb1bdd6 (last accessed 28/11/21).

Lipsitch, M. 2020, "Good Science Is Good Science", *European Journal of Epidemiology*, 35, 519-22.

Lohse, S. and Bschir, K. 2020, "The COVID-19 Pandemic: a Case for Epistemic Pluralism in Public Health Policy", *History and Philosophy of the Life Sciences*, 42, 4, 58.

Mazzocchi, F. 2021, "Drawing Lessons from the COVID-19 Pandemic: Science and Epistemic Humility Should go Together", *History and Philosophy of the Life Sciences*, 43, 3, 92.

Merton, R.K. 1942, "A Note on Science and Democracy", *Journal of Legal and Political Sociology*, 1, 115-26.

Meyer, R. and Madrigal, A.C. 2021, "Why the Pandemic Experts Failed: We're still Thinking about Pandemic Data in the Wrong Ways", *The Atlantic*, March 15, https://www.theatlantic.com/science/archive/2021/03/americas-coronavirus-catastrophe-began-with-data/618287/ (last accessed 07/11/21).

Monasterio Astobiza, A. 2021, "Science, Misinformation and Digital Technology during the Covid-19 Pandemic", *History and Philosophy of the Life Sciences*, 43, 2, 68.

Norris, J. 2020, "New COVID-19 'Citizen Science' Initiative Lets any Adult with a Smartphone Help to Fight Coronavirus", *UCFS news*, March 30, https://www.ucsf.edu/news/2020/03/417026/new-covid-19-citizen-science-initiative-lets-any-adult-smartphone-help-fight (last accessed 07/11/21).

Ongaro, M. 2021, "Making Policy Decisions under Plural Uncertainty: Responding to the COVID-19 Pandemic", *History and Philosophy of the Life Sciences*, 43, 2, 56.

Peckham, O. 2020, "Rosetta@home Rallies a Legion of Computers Against the Coronavirus", HPCwire, March 24, https://www.hpcwire.com/2020/03/24/rosetta-home-rallies-a-legion-of-computers-against-the-coronavirus/ (last accessed 07/11/21).

Reydon, T.A.C. 2020, "How Can Science Be Well-Ordered in Times of Crisis? Learning from the SARS-CoV-2 Pandemic", *History and Philosophy of the Life Sciences*, 42, 4, 53.