Participatory surveillance for COVID-19 trends detection in Brazil

Salome Wittwer

University of Zurich

Onicio Batista Leal Neto (✉ onicio@gmail.com)

University of Zurich

Daniela Paolotti

Institute for Scientific Interchange

Guilherme Lichand

University of Zurich

Research Article

Keywords:

Posted Date: March 16th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1247739/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Participatory surveillance for COVID-19 trends detection in Brazil

S. Wittwer¹; O.B. Leal-Neto^{1*}; D. Paolotti²; G. Lichand¹

¹Department of Economics - University of Zurich, Switzerland

²ISI Foundation, Italy

*Corresponding author

Abstract

The ongoing COVID-19 pandemic has emphasized the necessity of a well-functioning surveillance system to detect and mitigate disease outbreaks. Traditional surveillance (TS) usually relies on healthcare providers and generally suffers from reporting lags that prevent immediate response plans. Participatory surveillance (PS), an innovative digital approach whereby individuals voluntarily monitor and report on their own health status via Web-based surveys, has emerged in the past decade to complement traditional data collections approaches. This study compares novel PS data on COVID-19 infection rates across nine Brazilian cities with official TS data to examine the opportunities and challenges of using the former, and the potential advantages of combining the two approaches. We find that high participation rates are key for PS data to adequately mirror TS infection rates. Where participation was high, we document a significant trend correlation between lagged PS data and TS infection rates, suggesting that the former could be used for early detection. In our data, forecasting models integrating both approaches increased accuracy up to 3% relative to a 14-day forecast horizon model based exclusively on TS data. Furthermore, we show that the PS data captures a population that significantly differs from the traditional observation. These results corroborate previous studies when it comes to the benefits of an integrated and comprehensive surveillance system, but also shed lights on its limitations, and on the need for additional research to improve future implementations of PS platforms.

Introduction

The global COVID-19 pandemic in March 2020 has had unprecedented consequences around the world. It caused widespread illness and deaths, as well as worldwide economic, political and social repercussions¹. The occurrence of such an extraordinary event emphasizes the need of well-functioning disease surveillance systems to detect and monitor disease outbreaks and epidemics. Countries use disease monitoring systems to assess, predict and mitigate infectious disease outbreaks^{2,3}. Reliable and timely data are critical to protect populations and build the foundation for governments, policy makers and officials to intervene and prevent widespread infections⁴. As such, developing and improving on existing surveillance methods remains a rapidly growing and emerging field⁵.

Most current surveillance systems rely largely on traditional healthcare institutions, such as clinics, hospitals, and laboratories, to systematically collect data from practitioners as a public health monitoring tool². Health care providers send reports to public health officials with a certain regional or national data aggregation. In a few cases (usually, only a small percentage), these reports are then confirmed by laboratory analysis. Reported cases are then accounted as official disease cases⁶⁻⁸. Since the data is sourced from different institutions, aggregation often involves several time-lags throughout the chain of data collection, reducing the timeliness of measures and actions^{4,9,10}. Moreover, the true burden of disease is often underestimated. In industrialized countries, healthy adults with no previous conditions usually do not visit a doctor if their symptoms remain mild. In emerging countries, socio-economic weak communities lack access or financial resources to seek medical aid, and thus, are overlooked by the traditional surveillance (TS) system¹¹. Over the last two decades, new digital disease surveillance approaches have emerged to supplement traditional data collection, such as *Participatory Surveillance (PS)*. Participatory disease surveillance is understood as an approach that directly engages the public in providing health data. Individuals monitor and assess their own health status and are encouraged to submit self-reports through digital platforms using mobile apps, websites, and phone-based surveys via text messages (SMS) or automated calls (Interactive Voice Response units; IVR)¹². Any individual can register on the platform (if they live in a country where a PS platform is deployed) and participate on a

voluntary basis. As such, digital crowdsourced data can be aggregated and analyzed in large numbers^{13–15}. Users are asked to regularly complete a questionnaire aimed at gathering information about their current (lack of) symptoms, access to health care, risk exposure and medication. In the case of COVID-19, the main difference compared to TS is the lack of lab-confirmed testing: positive cases are categorized as such only based on the reporting of certain symptoms, with a so-called syndromic surveillance approach. The added value of PS systems is that they can reach even individuals that do not engage with health care providers due to financial or cultural reasons, because they live too remotely to have access to health care facilities, or because their symptoms are too mild to cause concern¹². The rapid global increase in the use of mobile phones and wide access to the internet is largely responsible for the rise of such digital surveillance systems³. By integrating an additional subset of the population not covered by TS, the complementary data provide an additional layer of surveillance, potentially enabling more accurate detection of ongoing infections as well as anticipation of trend changes^{13,14}.

Participatory surveillance systems have so far proven to be accurate and reliable for influenza-like-illness (ILI) surveillance^{15,16}. Most recently, as testing capacities were exhausted across many countries in the context of the pandemic, PS systems have been implemented to support traditional systems in monitoring and controlling COVID-19 infections^{17,18}. So far, Brazil is the first and only Latin American country that has implemented a PS system at large scale to carry out syndromic surveillance – specifically, during the 2014 FIFA World Cup¹⁹ and the 2016 Olympic Games²⁰. Lately, a local Brazilian health authority has used a PS platform to complement the traditional system, with the goal of optimizing the targeting of test areas during the COVID-19 pandemic¹⁷. This PS system has been shown to be beneficial in identifying risk clusters for infections in this context; in particular, it was able to cover blind spots of the TS system, showcasing the potential to increase its sensitivity by complementing it with additional data, and to allocate scarce resources more efficiently by prioritizing certain areas for the distribution of test kits. Nevertheless, health agencies and governments are still hesitant to utilize the innovative and digital PS approach, as it is not yet fully recognized as a complementary source of disease surveillance.

Low public funding and sub-optimal resource allocation persist in the Brazilian health sector²¹. The limitations of the traditional surveillance system stress the need to reduce the burden of disease among vulnerable and socioeconomic weak communities. This motivates the following study to examine the opportunities and challenges of a PS system in the context of COVID-19 case detection across nine Brazilian cities. During seven months of the global pandemic in 2020, several city-level governments implemented the *Brazil Sem Corona* PS platform to gather additional insights on the spread of the disease. We aim to investigate the capability of PS data to adequately mirror traditional infection rates captured through TS, as well as the relevance of citizen participation for the identification of trends in COVID-19 cases. Furthermore, we investigate the potential benefits of combining the PS and TS systems for forecasting case trends. Among the benefits, we show that the PS system captures a part of the population that is so far overseen by traditional sources.

Results

Sufficient participation needed for an adequate representation

To assess the importance of participation, we compare the capability of the PS data to mirror the infection rates reported by the TS systems implemented in the nine cities. In Table 1, we show the nine Brazilian cities along with their level of participation. As the population size varies between the cities, we additionally weight the number of submitted reports by the population size displayed by the variable ‘Number of Reports by 100’. As participants could report several times throughout the observation period, there is a relatively large number of submitted reports compared to the number of participants. Even though São Paulo recorded a relatively large number of submitted reports, it is ranked only second to last due to its large population size. Teresina, Caruaru and Santo André are ranked as the top three based on the variable ‘Number of Reports by 100’. The large participation in those three cities is not coincidental, but rather, driven by massive local media campaigns led by local officials^{22–24}. Those social media campaigns seem to have impacted the participation behavior underlining the government’s role. Additionally, we display the variable ‘Share of Zeros’, which captures the share of observation days on which no possible COVID-19 case was recorded on the platform within a city. This problem is often

referred to as *zero inflation* (as the sample contains an excess of zeros) and is equivalent to censored data. Those numbers are negatively correlated with the level of engagement, as the cities at the bottom of the ranking display a striking share of zeros, as high as 96%.

City	Number of Reports	Number of Reports per 100	Number of Participants	Share of Zero's
Teresina	27,558	3.17	4,449	7.0%
Caruaru	10,279	2.81	1,363	32.0%
Santo Andre	14,207	1.97	2,866	46.7%
Niteroi	6,757	1.31	897	62.9%
Recife	3,895	0.24	492	73.9%
Porto Alegre	2,779	0.19	336	96.1%
Campinas	2,006	0.17	307	92.9%
Sao Paulo	12,453	0.10	2,360	70.7%
Rio de Janeiro	3,071	0.05	734	74.6%

Table 1: Participation Ranking Table 1 shows the nine Brazilian cities under examination. The ranking is based on the variable 'Number of Reports per 100', from largest to smallest. This variable shows the number of submitted reports per 100 inhabitants and is calculated as Number of Submitted reports divided by population size times 100. The variable 'Share of Zeros' shows the share of observation days on which none of the submitted reports were classified as a potential COVID-19 case

Every submitted report is then categorized into either *negative COVID-19 case* or *suspected COVID-19 case* based on the symptoms reported in the survey. More details on the categorization process as well as on the calculation of the infection rates can be found in the Chapter Methods. In Figure 1, we plot both the TS and PS incidence rates over the entire seven-month observation period. This allows us to analyze differences in detected infection rates between the two surveillance methods. The nine cities are arranged according to the participation ranking in Table 1, from top-left to bottom-right. The top row displays the three cities with the largest community engagement: Teresina, Caruaru and Santo André. Here, infections identified through PS and TS follow a relatively similar pattern, at least between March and August. For most cities, the trend of detected cases diverges between the two methods towards the end of the observation period, what can be explained by the relatively low number of submitted reports over that interval. More details on the daily number of submitted reports can be found in SI Figure 1. From the graphical analysis, it follows that the capability of the PS system to mirror the trend of

infections seems to decrease with low participation rates. This finding is supported by Table 2, which reports the Pearson Correlation Coefficients between the PS and TS time series within each city. Table 2 confirms that the largest correlation between the two data sources is indeed found for the cities with

Figure 1: Daily Infections per City Figure 1 shows the daily PS and TS infection rates found within each city. The graphs are arranged according to the weighted participation from top left to bottom right.

the largest participation rates, namely Teresina, Caruaru, and Santo André. The Pearson Correlation coefficients for those three cities are above 0.5, indicating a moderate to high positive correlation between the two data sources. Indeed, newspapers and officials in these three cities have intensely promoted the *Brazil Sem Corona* platform^{22–24}. Online advertising has also likely incentivized many individuals to participate in the initiative, leading to better insights from the PS system.

However, in cities without advertising campaigns, it is striking that the PS system was not able to mirror traditional infection rates adequately. In Campinas, the city with the lowest total number of submitted reports, PS performed so poorly that its data displayed a *negative* correlation with TS data. With a total

of only 2,006 submitted reports over a period of 210 days (seen in Table 1), a daily average of self-reports just under ten was observed, implying that the sample size is far too small. The same is true for the other cities at the low-end of the ranking, with Pearson correlations below 0.25. The only exception is São Paulo, whose correlation was only smaller than those of the top three cities. Even though it is ranked as second to last with regard to the population weighted participation, the total number of submitted reports for São Paulo was relatively large, at 12,453 reports. This corresponds to an average of nearly 60 submissions per day, resulting in a Pearson Correlation coefficient of 0.32. All in all, low participation is considered one of the main limitations to the capability of PS data to mirror traditional infection rates.

City	Pearson Correlation	t-statistics	p-value
Teresina	0.82	20.87	<0.0000
Caruaru	0.79	18.60	<0.0000
Santo Andre	0.56	9.83	<0.0000
Niteroi	0.24	3.60	0.0004
Recife	0.23	3.48	0.0006
Porto Alegre	0.18	2.66	0.0084
Campinas	-0.34	-5.04	<0.0000
Sao Paulo	0.32	4.97	<0.0000
Rio de Janeiro	0.09	1.30	0.1939

Table 2: Pearson Correlations Table 2 shows the Pearson Correlation coefficients for each city along with the t-statistics and p-values. The cities are ranked according to the population weighted participation from largest to smallest.

Above and beyond low participation, as previously indicated, Campinas and other cities at the bottom of the ranking likely suffer from a zero inflation problem, with up to 96% of the submitted self-reports indicating minor to no symptoms. Along with low participation, such a significant share of reports being categorized as negative COVID-19 cases seems to make an adequate representation of the traditional data extremely difficult and limits its benefits. To reap its maximal benefits, it must be ensured that the population is well informed about the existence of the PS system, and that individuals can be sufficiently motivated to report accurately regardless of their symptoms.

Studies conducted in the US and Western Europe suggest that the population sub-group captured by the *Brazil Sem Corona* PS system does not necessarily represent the general Brazilian population²⁵⁻²⁷. In previous PS data, female participants were significantly over-represented compared to the general population. Besides that, age groups below 30 and above 80 were under-represented. Moreover, the average participant most likely holds a higher educational degree than the average population²⁸. Also, individuals living in bigger cities are more likely to participate, leading to clusters around more urban areas and information gaps in more remote regions. This is supported by the geo-coordinates from the submitted reports on the *Brazil Sem Corona* platform, which indicate that a significant share of participants lives in or around urban areas. Targeting the under-represented population may prospectively improve the complementary benefit of the platform, particularly among those not seeking medical attention. Despite the biases found in the population covered by PS, studies highlight that PS systems that engage a sufficiently large group of participants can still adequately capture TS infection trends^{27,29}.

We emphasize the successful collaboration between governments, locals, and professionals when it comes to maximizing the use and gain of the *Brazil Sem Corona* platform. The three cities with extensive social media campaigns clearly demonstrate the capabilities of the PS system - under the condition of sufficient participation. Defining sufficient participation remains relevant for future research; so far, previous studies only highlight the number of reports as a critical element and claim the need to maintain sufficient coverage without setting a certain threshold^{13,26,30}. Factors such as population density, urbanization and area size most likely cause regional variations, making it significantly more difficult to determine a specific number that is generally valid.

Next, we analyze the participation behavior of the individuals across all nine cities. A significant share of 40% - 50% of individuals participated only once throughout the observation period, while roughly 6% - 13% are considered as frequent participants, with at least 15 submitted reports. Detailed numbers can be found in the SI Table 2. Ideally, participatory surveillance tracks individuals throughout the season. Knowing that nearly half of the participants did not continue their engagement after the first

report submission stresses the need for efforts to ensure more frequent participation in the future. Previous studies from the US and Canada have found significant differences in participation across age groups, along with a 25% lower probability of frequent participation by women relative to men²⁷.

PS time lags increase PC coefficients

Another important aspect of complementing disease surveillance systems is the goal of early trend detection, to identify disease outbreaks. Integrating multiple data sources not only aims at improving data insights and detecting more cases but, ideally, leads to the detection of outbreaks at an earlier stage. We recalculate the Pearson correlation coefficients for the three key cities with the largest public engagement using a seven-day and 14-day lag in the PS data series. Panel A in Table 3 shows the Lagged Pearson Correlations for the full observation period, and Panel B presents the results for the reduced period from April 1st to July 31st.

In Panel A, an increased correlation can only be found for Teresina, where the coefficient increases from 0.82 to 0.89 for the 14-day lag. For Caruaru and Santo André, the coefficients decline from 0.79 to 0.71 and from 0.56 to 0.45, respectively. However, the correlations for the seven-day lag remain on roughly the same level. While these results are ambiguous and rather weak, the picture gets clearer in Panel B.

A: Full Observation Period				
City	PC	PC 7-day Lag	PC 14-day Lag	Avg. Daily Reports
Teresina	0.82	0.87	0.89	128
Caruaru	0.79	0.77	0.71	49
Santo Andre	0.56	0.56	0.45	66

B: Reduced Observation Period				
City	PC	PC 7-day Lag	PC 14-day Lag	Avg. Daily Reports
Teresina	0.86	0.91	0.93	193
Caruaru	0.85	0.85	0.88	76
Santo Andre	0.82	0.89	0.88	88

Table 3: Lagged Pearson Correlations A: Full Observation Period displays the PC values for the full seven-month period. B: Reduced Observation Period displays the PC values for the reduced period between April and July.

In Panel B, the coefficients remain either constant or increase using a seven-day and 14-day lag across all three cities. By removing the tails of the observation period, the focus is set on the months with the

strongest engagement, as it is shown by the higher daily average of submitted reports in Panel B. The finding of greater coefficients in the Lagged Pearson correlation indicates that the timeliness of the PS system helps identifying slightly preceding trends. This supports the idea that a PS system which engages a larger volunteer network is more likely to mirror infection trends, resulting in better data insights, and, eventually, in earlier anticipation of trend changes. Recognizing outbreak patterns only slightly in advance might already have great benefits for health agencies when it comes to fighting pandemics.

PS integrates an additional subset of the population

Table 4 describes important characteristics found in the reports submitted for participants in Teresina, Caruaru and Santo André. It shows that the population which is captured through the *Brazil Sem Corona* platform at least partly differentiates from the one represented by the traditional system.

City	Share of Confirmed Cases	Share of Health-Care Seekers	Share of Medication Intakers
Teresina	36.7%	27.4%	50.2%
Caruaru	42.7%	31.4%	52.0%
Santo Andre	57.1%	34.6%	58.0%

Table 4: Characteristics of Potential COVID-19 case reports Table 4 shows the Share of Confirmed Cases, the Share of Health-Care Seekers and the Share of Medication Intakers among all submitted reports that were categorized as Potential COVID-19 case.

In the traditional system, the new recorded COVID-19 cases per day are aggregated based on positive lab-confirmed tests. In contrast, the PS data shows a significant share of reports categorized as potential COVID-19 case that are not lab-confirmed. Teresina, the city with the greatest public engagement, featured the lowest share of confirmed cases, only 36.7%. While this share is also covered by the traditional surveillance system, the remaining 63% of the reports are not, as they are only categorized based on the reported symptoms without ever being officially tested. With decreasing participation, the numbers of confirmed cases seem to increase for the city of Caruaru and Santo André, with shares of 42.7% and 57.1% respectively. This result highlights the benefits of a PS implementation, as it allows

to complement the population captured by traditional sources. The greater the engagement, the larger is the impact of a PS system as it maximizes the inclusion of individuals overseen by traditional surveillance. While, for Teresina, 63% of the cases detected by the PS system are not covered by traditional sources, the share goes down to 33% for Santo André. The information on medication intake and healthcare seeking behavior supports this statement. Only around one third of the reports indicate a visit to a health-care facility, and slightly more than half report medication intake. The negative correlation with participation spans all the variables. Amidst a pandemic, when resources are scarce and testing and care capabilities are limited, it can be of great importance to detect additional infections among different subgroups of a population.

14-day forecast accuracy slightly improved by PS integration

For the three cities with an at least moderate positive Pearson correlation coefficients, we compare the RMSEs and MAEs from the *Baseline model* with the *Combination* as well as the *Lagged Combination model* (Table 5). Details on each model can be found in the Methods Chapter. We present a horizon of one-day, seven-day and 14-day ahead forecast. The errors for the one-day ahead forecasts are generally smaller than the errors found in the seven- and 14-day horizon for all three cities. This is not surprising, as uncertainty rises with longer forecast horizons and hence, the forecast accuracy of a model is lower. When looking at the one-day forecasts, the Baseline model seems to outperform the Combination model, at least for Teresina and Santo André. As mentioned before, the uncertainty grows with longer horizons, which might explain the lack of value added that comes from integrating PS data in a one-day forecast. In Teresina, however, both the seven-day and 14-day forecast from the Combination and Lagged Combination models perform better than the Baseline model, with slightly lower RMSEs. The Combination model shows a 4.2% reduction in the RMSE for the seven-day horizon and a 2.6% reduction for the 14-day horizon. The Lagged Combination model shows reductions of 0.2% and 2.1% respectively. In Caruaru, the Combination model performs slightly worse for the seven-day forecast relative to the Baseline model, with a 0.5% increase in the RMSE, but improves the forecast in the 14-day horizon, with a 1.7% reduction in the RMSE. Similar results are found for the Lagged Combination

model, whereby the seven-day horizon forecast shows an 4.1% increase in RMSE, while the 14-day forecast improves it by 0.4%. The Combination model applied on the data from Santo André again reduces the RMSEs for both the seven-day and 14-day forecast, by 2.8% and 2.7% respectively. However, the Lagged Combination model is not able to improve the forecast for any of the applied horizons.

A: Teresina			
Forecast period	Model	RMSE	MAE
Forecast t+1	Baseline model	0.0195	0.0133
	Combination model	0.0198	0.0140
	Lagged Combination model	0.0193	0.0137
Forecast t+7	Baseline model	0.0410	0.0302
	Combination model	0.0393	0.0281
	Lagged Combination model	0.0409	0.0310
Forecast t+14	Baseline model	0.0347	0.0245
	Combination model	0.0338	0.0246
	Lagged Combination model	0.0340	0.0249
Note: The models used $n=13$ lagged components as independent variables.			
B: Caruaru			
Forecast period	Model	RMSE	MAE
Forecast t+1	Baseline model	0.0214	0.0162
	Combination model	0.0212	0.0160
	Lagged Combination model	0.0290	0.0220
Forecast t+7	Baseline model	0.0316	0.0234
	Combination model	0.0318	0.0237
	Lagged Combination model	0.0329	0.0244
Forecast t+14	Baseline model	0.0293	0.0190
	Combination model	0.0288	0.0187
	Lagged Combination model	0.0292	0.0190
Note: The models used $n=5$ lagged components as independent variables.			
C: Santo Andre			
Forecast period	Model	RMSE	MAE
Forecast t+1	Baseline model	0.0137	0.0094
	Combination model	0.0236	0.0181
	Lagged Combination model	0.0203	0.0157
Forecast t+7	Baseline model	0.0225	0.0169
	Combination model	0.0219	0.0174
	Lagged Combination model	0.0321	0.0245
Forecast t+14	Baseline model	0.0292	0.0245
	Combination model	0.0284	0.0236
	Lagged Combination model	0.0299	0.0251
Note: The models used $n=14$ lagged components as independent variables.			

Table 5: Forecasting errors A: Displays the forecasting errors for the city of Teresina using a Baseline model, Combination model and Lagged Combination model. Errors are calculated for a one-day, seven-day and 14-day horizon. The models integrate an optimal number of lagged components equal $n=13$. B: Similar to Panel A, it shows the results for the city of Caruaru. The models integrate an optimal number of lagged components equal $n=5$. C: Similar to Panel A, it shows the results for the city of Santo Andre. The models integrate an optimal number of lagged components equal $n=14$.

Even though improvements are only modest, there is a pattern that can be identified across all three cities. The Combination model outperforms the Baseline model for the 14-day forecasts by up to 2.7%. The results for the seven-day forecasts are more ambiguous: while there are improvements of up to 4.1%, only two out of three cities show reduced RMSEs. The results indicate that, under greater uncertainty – which is usually found in longer forecast horizons –, a participatory surveillance system has the ability to improve forecast accuracy by complementing traditional data. One reason we see for the relatively small magnitude of improvements may be the adjustment in government behavior. Because of the severity of the COVID-19 pandemic, governments and health authorities were forced to overcome administrative burdens and provided traditional data *on a daily basis*. As a result, the advantage of timeliness that usually distinguishes a PS system from traditional surveillance pales under the circumstances.

Even if the RMSEs improve only minimally or remain on roughly the same level, integrating both sources still brings a benefit. As it is shown in Table 4, up to 66% of the examined reports categorized as possible COVID-19 cases are not confirmed, and therefore, not captured by the TS system. When complementing one another, an additional subset of the population is taken into account – with the potential to enhance the early anticipation of trend changes and health threats across a larger and more diverse population. Once again, the larger the participation within a city, the better the representation of the population and the more valuable is the integration of a PS system. With the additional signal that is captured by the PS data, health authorities can allocate resources and services more efficiently. Improvements in 14-day forecasts allow health officials to respond more quickly and prioritize certain areas identified as more likely to suffer from rising infection numbers. This may be of particular importance for low-income countries, or for regions that suffer from a considerable scarcity in health services. The PS system brings value by producing information that can be used to reduce uncertainty in allocation decisions. Furthermore, infection monitoring can be improved thanks to the geo-location information provided by the PS data, which allows for a typically higher spatial resolution. Aside from preventing local transmissions, improvements in the surveillance system may result in externalities, such as lower infections in other regions³¹.

Discussion

For the PS system in Brazilian cities, the analysis concludes that participation is highly relevant for adequate data insights. We find that the PS infection rates from the three cities with largest participation adequately mirrored the TS infection rates, even though the representativeness of the PS sub-population is most likely biased in terms of sex and age. These three cities were able to engage a large group of individuals due to social media campaigns promoted by local officials and governments. The insights from the other cities are far less valuable as the data is not able to represent the traditional infection rates due to low engagement and zero inflation. Local governments and health authorities must address the underrepresentation of certain socio-demographic characteristics of the population and engage a sufficiently large group of volunteers to maximize the benefits of a PS system. Complementing traditional disease surveillance systems may further increase the possibility for early identification of outbreaks under the condition of a sufficiently large participation. Slight improvements in the forecasting accuracy of up to 3% are identified for models integrating data from both surveillance sources compared to Baseline models relying entirely on traditional data. Even if forecasting improvements are only weak, the detection of infections is improved as PS can cover an additional subset of the population that is overseen by the traditional system. The above findings contribute to a deeper understanding on the benefits of a complementary digital surveillance layer. They corroborate previous literature, emphasizing that the two approaches can be complements for timely health threat identification^{3,14,16}. Even though some of the results indicate only small improvements in accuracy, the possibility of enhancing case detection through broader coverage cannot be neglected.

We hypothesize that severity of COVID-19 likely influenced individuals' motivation to engage in a voluntary surveillance system. Governments have adjusted their behavior during the ongoing pandemic and made significant efforts to improve the timeliness of traditional data reporting. Many countries, including Brazil, have overcome administrative burdens and were ultimately able to report daily new infections. As such, the timeliness that usually distinguishes a PS system from traditional surveillance has most likely vanished. We see this as a reasonable explanation for why the magnitude of the results

from the forecasting results is weaker compared to previous studies conducted in the field of ILI^{14,15}. Besides that, ILI tracking allows for a data collection across several flu seasons, increasing the amount of available data and allowing for consistency checks across seasons, whereas the novelty of COVID-19 only allows for a single seven-month observation period in Brazil.

The implementation of a proper surveillance system remains an important challenge in developing countries. While the PS system aims to address disparities in health outcomes, it currently reaches more urban regions and relatively more educated people. Therefore, voluntary crowd-sourced data likely contains a population bias. Assessing the potential benefits for more rural communities remains open to future research in order to realize the system's full potential. Engaging and motivating a greater diversity of individuals remains a key challenge that must be addressed in future PS platform implementations. Obstacles such as lack of access to modern technologies, illiteracy or simply lack of awareness of the benefits from participation might hinder progress towards a more diverse reporting population.

Nearly half of the participants in Brazil submitted only a single report. This prevents monitoring individuals' health status over a longer observation period and, as a result, reduces the insights that can be gained from PS systems. This stresses the need to evaluate incentives to induce more frequent participation. Additionally, in-depth research is needed to determine the reasons that lead to a discontinuation of participation after the first submission. Addressing the above issues in future PS platform implementations is likely to lead to even greater benefits.

The unavailability of data on participants' socio-demographic characteristics in the *Brasil Sem Corona* platform prevents an important analysis on the representativeness of participants versus the general Brazilian population. Performing further tests and extending the method could confirm the validity of the results and deepen the insights. The forecasting models were kept rather simple on purpose, with only the TS and PS data as independent variables. Adding mobility data or contact tracing data, for instance, could improve accuracy, leading to a better understanding of transmissions. There are further limitations to keep in mind, such as the impact of the categorization criteria of a submitted report into either *suspected COVID-19* case or *negative case*, as it influences the number of reports included in the

case counts and the likelihood of false positives. Furthermore, the applied local regression ('LOESS') remains an approximation to smooth the highly volatile data. Alternative smoothing techniques might produce different results, which could lead to divergent interpretations.

Participatory surveillance has not yet been fully recognized as a complementary surveillance source. Future research needs to identify determinants of participation and proper incentives to induce larger coverage and higher diversity of participants. As the quality of data insights improve, PS benefits may further expand. Deeper insights allow for greater acceptance and credibility among governments and health authorities. Expanding collaboration between researchers, officials and health authorities is needed to leverage data insights into timely response plans, which ultimately lead to better health outcomes.

Quantifying the economic value of a PS system implementation remains hard. But scarce public funds as well as persisting constraints to the TS system motivate for a PS system, making it an important avenue for future research. The decision to set up a PS system requires careful evaluation of its expected benefits, relative to the costs of setting up platforms and incentivizing engagement to increase both coverage and consistent reporting over time. The ability to compute such economic trade-offs might be key to have PS become a more integral part of policy toolkits moving forward.

Method

To assess the complementary benefit of a PS system in Brazil, we compare daily official COVID-19 infections on a municipality level with daily PS infection numbers. Below we describe the data used in the study as well as the methods we apply.

Traditional Surveillance Data

The traditional surveillance data for Brazil, prospectively called the TS data, is publicly accessible on GitHub³². It aggregates the official lab-confirmed daily new COVID-19 cases on a municipality level. Using the 2020 population size estimates for each city³³, we calculate the infection rate per 100'000 inhabitants the following way:

$$TS \text{ Infection rate}_{i,t} = \frac{\text{Number of new cases}_{i,t}}{\text{Population size}_i} * 100,000, \text{ where } i \text{ denotes the city and } t \text{ the day.}$$

Participatory Surveillance Data

The participatory surveillance data, further referenced as PS data, was collected through the *Brazil Sem Corona* platform³⁴. To gather information on an individual's health status, each participant was asked to fill out a questionnaire on symptoms as well as exposure. Before filling out this questionnaire, participants were asked to agree with an informed consent form in the registration, describing the study and the purpose of the project. The list of symptoms was based on the COVID-19 case definition and contained the following: fever, cough, shortness of breath, runny nose, sore throat, headache, fatigue, nausea, rash, joint pain, chills, diarrhea and loss of taste. Additionally, they were asked about medication intake and whether they sought a healthcare facility. It was requested to Colab the access to the anonymized dataset of *Brazil Sem Corona*. The access to the data and study was approved by the Colab Institutional Management Board. All methods were carried out in accordance with the guidelines and regulations, including, but not limited to, the *Lei Geral da Proteção de Dados – LGPD*, the official regulation on data privacy and protection valid in Brazil.

The platform was set online on March 20th, right after the WHO officially characterized Covid-19 as a pandemic on March 11³⁵. Brazilian newspapers and magazines started to advertise the use of the application causing an increase in self-reports submissions²²⁻²⁴. Reports were only collected between March 20th and October 20th, limiting the analysis to this seven-month observation period. Even though the platform was accessible all-around Brazil, nearly 65% of the reports were submitted from nine cities. Therefore, the study focuses only on those cities with the largest number of submitted reports. People from Teresina, Caruaru, Santo Andre, Niteroi, Recife, Porto Alegre, Campinas, Sao Paulo and Rio de Janeiro submitted an aggregated total of 83005 reports from 13582 individuals. Based on the symptoms reported in the questionnaire, each report is then categorized into either no symptom, light symptom, suspected COVID-19 case, severely suspected COVID-19 case or confirmed case. In order to be categorized as a suspected case, the user must report fever together with at least one other symptom. If along these symptoms either a medication intake is reported or a health care facility is sought, it is elevated to severely suspected. The reasoning behind this is an indication of stronger symptoms that can be related to the severity of a possible infection. Categorized as confirmed case are only those that report a positive lab-tested COVID-19 test result. All the suspected, severely suspected and confirmed cases are then treated as "*possible COVID-19 cases*", while light and no symptom reports are treated as "*negative cases*". The inclusion of not only confirmed cases but also suspected cases is one of the key differences to traditional case counts. For each city individually, we aggregate the reports submitted on the same day and calculate a daily infection rate:

$$PS\ Infection\ rate_{i,t} = \frac{Number\ of\ possible\ cases_{i,t}}{Number\ of\ submitted\ reports_{i,t}}, \text{ where } i \text{ denotes the city and } t \text{ the day.}$$

Data volatility is addressed by applying a simple but powerful tool called LOESS. It fits smooth lines to empirical data using a non-parametric approach³⁶. For all nine cities, the smoothed PS and TS infection rates are plotted in Figure 1.

Pearson Correlation Calculation

To measure the statistical relationship between the two series, we calculate the Pearson Correlations (PC). The PC coefficient provides information on both the direction of the relationship and on the magnitude. Besides calculating the PC_t coefficients, we additionally use a seven as well as a 14-day lagged PS series to calculate PC_{t-7} and PC_{t-14} . A stronger correlation between the TS and lagged PS time series supports the theory of early trend detection which is one goal pursued by participatory surveillance. However, this is only conducted for three cities which show a strong positive PC_t above at least 0.5. It is found that towards the end of the observation period only few people participated on the platform, such that around 80% of the reports were submitted within the period between April 1st to July 31st. Besides calculating the lagged Pearson Correlations, we recalculate the coefficients for a reduced four-month observation period during which engagement was largest. Panel A in Table 4 shows the Lagged Pearson Correlations for the full observation period while Panel B presents the results for the reduced period.

Forecasting models

Lastly, we measure the value-added stemming from the implementation of a PS system using three forecasting models to predict future incidence rates. For this study, the traditional data always represents the ground truth. The first model, further referenced as *Baseline model*, is a univariate model based only on TS data. The second and third model, called *Combination model* and *Lagged Combination model*, are bivariate models integrating both TS and PS data. However, the third model uses a 14-day lag in the PS incidence rate series. For all three models, we use a linear auto-regressive model with n daily lagged components. For each city, the optimal number of independent variables is selected based on the AIC. Thereby, the explained part of the variation is maximized while using only the lowest possible amount of time lags. An overview of the models can be found below:

Baseline models	Combination models	Lagged Combination models
$\widehat{TS}_{i,t+1} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l}$	$\widehat{TS}_{i,t+1} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l} + \sum_{l=0}^n \widehat{\beta}_l PS_{t-l}$	$\widehat{TS}_{i,t+1} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l} + \sum_{l=14}^n \widehat{\beta}_l PS_{t-l}$
$\widehat{TS}_{i,t+7} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l}$	$\widehat{TS}_{i,t+7} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l} + \sum_{l=0}^n \widehat{\beta}_l PS_{t-l}$	$\widehat{TS}_{i,t+7} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l} + \sum_{l=14}^n \widehat{\beta}_l PS_{t-l}$
$\widehat{TS}_{i,t+14} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l}$	$\widehat{TS}_{i,t+14} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l} + \sum_{l=0}^n \widehat{\beta}_l PS_{t-l}$	$\widehat{TS}_{i,t+14} = \sum_{l=0}^n \widehat{\alpha}_l TS_{t-l} + \sum_{l=14}^n \widehat{\beta}_l PS_{t-l}$

Here, $\widehat{TS}$ stands for the estimation, where as TS stands for the true values. By weighing n past components, a one, seven and 14-day ahead forecast is estimated. The model parameters $\hat{\alpha}_l$ and $\hat{\beta}_l$ are estimated based on a training sub-sample which contains the first 80% of the data. The out-of-sample forecasting accuracy is then calculated based on the remaining 20% of the data. So, the first 80% of the observation period is used to fit the models while the remaining 20% is used to evaluate its performance.

To evaluate the forecasting models, we present common metrics such as the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

- The RMSE measures the difference between predicted and true values:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\widehat{TS} - TS)^2}$$

- The MAE measures the average of the absolute errors between predicted and true values:

$$MAE = \frac{1}{n} \sum_{i=1}^n |TS - \widehat{TS}|$$

While both evaluation metrics present an average prediction error, the RMSE weighs large errors relatively higher. Therefore, a focus is set on the RMSE for the interpretation of the results. The comparison of the prediction errors across the three models allows drawing a conclusion on the value-added stemming from the complementary data inclusion.

List of Figures and Tables

Figure 1: Daily Infections per City

Table 1: Participation Ranking

Table 2: Pearson Correlations

Table 3: Lagged Pearson Correlations

Table 4: Characteristics of Potential COVID-19 case reports

Table 5: Forecasting errors

Competing interests

The authors report no declarations of interest.

Data availability

The data that support the findings of this study are available from COLAB but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of COLAB.

References

1. Ingram, A. Outbreak, epidemic, pandemic: The politics of global health events: A review of Sara E. Davies' *Containing Contagion*. *Singapore Journal of Tropical Geography* **42**, 350–352 (2016).
2. Mostashari, F. & Hartman, J. Syndromic Surveillance: A Local Perspective. *Journal of Urban Health* **80** (2003).
3. Salathé, M. Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health. *The Journal of Infectious Diseases* **214**, 399–403; 10.1093/infdis/jiw281 (2016).
4. Smith, P. F., Hadler, J., Stanbury, M., Rolfs, R. & Hopkins, R. Blueprint version 2.0: updating public health surveillance for the 21st century. *Journal of public health management and practice : JPHMP* **19**, 231–239 (2013).

5. Abat, C., Chaudet, H., Rolain, J.-M., Colson, P. & Raoult, D. Traditional and syndromic surveillance of infectious diseases and pathogens. *International Journal of Infectious Diseases* **48**, 22–28; 10.1016/j.ijid.2016.04.021 (2016).
6. CDC. Event-based Surveillance. Available at <https://www.cdc.gov/globalhealth/healthprotection/gddopscenter/how.html> (2019).
7. Institute of Medicine. *Managed Care Systems and Emerging Infections: Challenges and Opportunities for Strengthening Surveillance, Research, and Prevention: Workshop Summary* (The National Academies Press, Washington, DC, 2000).
8. Starnini, M., Aleta, A., Tizzoni, M. & Moreno, Y. Impact of data accuracy on the evaluation of COVID-19 mitigation policies. *Data & Policy* **3**, e28; 10.1017/dap.2021.25 (2021).
9. Nsubuga, P., White, M. E. & Thacker, S. B. Public Health Surveillance: A Tool for Targeting and Monitoring Interventions. 53. In *Disease Control Priorities in Developing Countries, 2nd edition*, edited by Dean T. Jamison, *et al.* (The International Bank for Reconstruction and Development / The World Bank, Washington DC, 2006).
10. Ramalingam, B. Real-time Monitoring in Disease Outbreaks: Strengths, Weaknesses and Future Potential. *Institute of Development Studies* (2016).
11. Peters, D. H. *et al.* Poverty and access to health care in developing countries. *Annals of the New York Academy of Sciences* **1136**, 161–171; 10.1196/annals.1425.011 (2008).
12. Leal-Neto, O. B. *et al.* Combining Wearable Devices and Mobile Surveys to Study Child and Youth Development in Malawi: Implementation Study of a Multimodal Approach. *JMIR Public Health Surveill* **7**, e23154; 10.2196/23154 (2021).
13. Smolinski, M. S., Crawley, A. W., Olsen, J. M., Jayaraman, T. & Libel, M. Participatory Disease Surveillance: Engaging Communities Directly in Reporting, Monitoring, and Responding to Health Threats. *JMIR Public Health Surveill* **3** (2017).

14. Guerri, C. *et al.* Participatory Syndromic Surveillance of Influenza in Europe. *The Journal of Infectious Diseases* **214**; 10.1093/infdis/jiw280 (2016).
15. Perrotta, D., Tizzoni, M. & Paolotti, D. Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy. *Proceedings of the 26th International Conference on World Wide Web* (2017) % Reference for Method).
16. D. Paolotti *et al.* Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection* **20**, 17–21; 10.1111/1469-0691.12477 (2014).
17. Leal-Neto, O. B., Santos, F., Lee, J. Y., Albuquerque, J. O. & Souza, W. V. Prioritizing COVID-19 tests based on participatory surveillance and spatial scanning. *International Journal of Medical Informatics* **143**, 104263; 10.1016/j.ijmedinf.2020.104263 (2020).
18. Mahmud, A. S. *et al.* Participatory syndromic surveillance as a tool for tracking COVID-19 in Bangladesh. *Epidemics* **35**; 10.1016/j.epidem.2021.100462 (2021).
19. Leal-Neto, O. B. *et al.* Saúde na Copa: The World's First Application of Participatory Surveillance for a Mass Gathering at FIFA World Cup 2014, Brazil. *JMIR Public Health and Surveillance* **3**, e26; 10.2196/publichealth.7313 (2017).
20. Leal-Neto, O. B. *et al.* Participatory Surveillance Based on Crowdsourcing During the Rio 2016 Olympic Games Using the Guardians of Health Platform: Descriptive Study. *JMIR Public Health and Surveillance* **6** (2020).
21. Massuda, A., Hone, T., Leles, F. A. G., Castro, M. C. de & Atun, R. The Brazilian health system at crossroads: progress, crisis and resilience. *BMJ Global Health* **3** (2018).
22. CNN Brazil. Apple and Google team up to warn of COVID-19 contagion. Available at <https://www.cnnbrasil.com.br/tecnologia/apple-e-google-se-unem-para-alertar-contagio-de-covid-19/> (2020).

23. Época Negócios. Startups launch application to monitor risks of coronavirus infection in Brazil. Available at <https://epocanegocios.globo.com/Empresa/noticia/2020/03/startups-lancam-aplicativo-para-monitorar-riscos-de-infeccao-por-coronavirus-no-brasil.html> (2020).
24. Fórum de Saúde Digital. With the collaboration of citizens, the Colab application will detect the risk of coronavirus cases throughout Brazil. Available at <https://forumsaudedigital.com.br/com-colaboracao-de-cidadaos-aplicativo-colab-vai-detectar-risco-de-casos-de-coronavirus-em-todo-o-brasil/> (2020).
25. Cantarelli, P. *et al.* The representativeness of a European multi-center network for influenza-like-illness participatory surveillance. *BMC Public Health* **14** (2014).
26. Debin, M. *et al.* Evaluating the Feasibility and Participants' Representativeness of an Online Nationwide Surveillance System for Influenza in France. *PLOS ONE* **8**, 1–11; 10.1371/journal.pone.0073675.
27. Baltrusaitis, K. *et al.* Determinants of Participants' Follow-Up and Characterization of Representativeness in Flu Near You, A Participatory Disease Surveillance System. *JMIR Public Health Surveill* **3**, e18; 10.2196/publichealth.7304.
28. Bajardi, P. *et al.* Determinants of Follow-Up Participation in the Internet-Based European Influenza Surveillance Platform Influenzanet. *Journal of Medical Internet Research* **16**; 10.2196/jmir.3010 (2014).
29. Santillana, M. *et al.* Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology* (2015).
30. Wójcik, O. P., Brownstein, J. S., Chunara, R. & Johansson, M. A. Public health for the people: participatory infectious disease surveillance in the digital age. *Emerging Themes in Epidemiology* **11** (2014).
31. Vries, L. de, Koopmans, M., Morton, A. & van Baal, P. The economics of improving global infectious disease surveillance. *BMJ Global Health % References from Outlook % References from Conclusion* **6**; 10.1136/bmjgh-2021-006597 (2021).

32. Cota, W. Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level. *SciELOPreprints*:362; 10.1590/scielopreprints.362 (2020).
33. Instituto Brasileiro de Geografia e Estatística. Cities and States. Available at <https://www.ibge.gov.br/en/cities-and-states.html?view=municipio> (2020).
34. Brasil Sem Corona. Brasil Sem Corona Participatory Surveillance Platform. Available at <https://www.brasilsemcorona.com.br/> (2021).
35. WHO. Rolling updates on coronavirus disease (COVID-19). Available at <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen> (2020).
36. Jacoby, W. G. Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies* **19**, 577–613; 10.1016/S0261-3794(99)00028-1 (2000).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [S11.pdf](#)