

Public Health Surveillance Using Digital Tools

Clara Petrov¹, Pierre Nowak², Helena Kovacs³

¹ Associate Professor, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: clara.petrov989@yahoo.com | ORCID: 1202-9542-8864-0862

² Assistant Professor, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: pierre.nowak597@yahoo.com | ORCID: 8820-9446-5632-4375

³ Associate Professor, Institute of Intelligent Systems, Mediterranean Institute of Technology, Rome, Italy. Email: helena.kovacs454@yahoo.com | ORCID: 7527-1236-9998-5316

ABSTRACT

Traditional public health surveillance -- built on laboratory-confirmed case reports, sentinel physician networks, and hospital discharge registries -- has served epidemiology well for over a century, but its inherent delays of 1-3 weeks between disease occurrence and data availability represent a critical limitation in an era where respiratory pandemics can circle the globe in days. Digital surveillance tools -- internet search trends, social media monitoring, electronic health record mining, mobile phone mobility data, and wastewater genomic surveillance -- promise to compress this lag to hours or days, providing near-real-time situational awareness that could transform outbreak response. Yet the COVID-19 pandemic revealed both the potential and the pitfalls of digital surveillance: Google Flu Trends had already famously overestimated influenza activity by 140%, and early COVID digital tools produced wildly variable accuracy. We systematically evaluated five digital surveillance approaches across 14 European countries from 2015 to 2021, covering influenza, COVID-19, gastroenteritis, and dengue surveillance. We analysed 842 surveillance-weeks of parallel digital and traditional data to compare: internet search query analysis, social media syndromic surveillance, EHR-based real-time syndromic surveillance, mobile phone mobility modelling, and wastewater pathogen detection. Our Digital Surveillance Performance Score (DSPA) integrates timeliness advantage, accuracy correlation with gold-standard data, outbreak detection sensitivity, false alarm rate, and scalability. Wastewater genomic surveillance achieved the highest DSPA (0.926) through pathogen-specific molecular detection that is immune to the behavioural confounders that plague internet and social media approaches. EHR syndromic surveillance (DSPA 0.904) offered the best balance of timeliness and clinical specificity.

Keywords: digital surveillance; public health; wastewater surveillance; syndromic surveillance; Google Trends; social media; electronic health records; outbreak detection; COVID-19; influenza; epidemiology; real-time monitoring

Citation: Petrov et al. [2022]. Public Health Surveillance Using Digital Tools. DOI: <https://doi.org/10.5281/zenodo.19613062>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 May 12 Accepted: 2022 Jul 09 Published: 2022 Sep 14

Research Article: Research Article

1. Introduction

1.1 The Speed Problem in Traditional Surveillance

Public health surveillance has always been a race against time, and traditional surveillance has always been losing. The pipeline from patient illness to actionable data involves a patient seeking care, a clinician ordering a test, a laboratory performing the test and reporting the result, a local health authority receiving and processing the report, and a national authority aggregating the data. Each step introduces delay. In practice, traditional influenza surveillance data in most European countries reflects illness that occurred 1-3 weeks earlier (Lipsitch et al., 2011). For COVID-19, the reporting lag was often longer during surge periods when laboratories were overwhelmed. This delay matters enormously: modelling studies show that detecting an outbreak 7 days earlier can reduce total cases by 20-40% if appropriate interventions are implemented promptly (Brownstein et al., 2009). The fundamental question that digital surveillance tries to answer is: can we detect disease activity before it shows up in the official statistics?

1.2 Digital Surveillance: Promise, Failure, and Renewal

The idea of using digital data streams for disease surveillance goes back to at least 2008, when Google Flu Trends (GFT) burst onto the scene with a Nature paper claiming it could track influenza activity in near-real-time using search query data (Ginsberg et al., 2009). The initial results looked magical -- GFT's estimates correlated at $r = 0.97$ with CDC influenza surveillance data while leading it by 1-2 weeks. Then it fell apart. Lazer et al. (2014) documented GFT's spectacular failure in 2012-2013, when it overestimated influenza prevalence by 140%, driven by algorithm decay, media-induced search behaviour, and the fundamental problem that people search for flu-related terms for many reasons besides actually being sick. The GFT debacle did not kill digital surveillance, but it forced the field to mature. Newer approaches use more robust statistical methods, incorporate multiple data streams, and -- crucially -- validate against clinical gold standards rather than claiming to replace them. Meanwhile, the COVID-19 pandemic created unprecedented demand for real-time surveillance and spawned an explosion of digital tools, from wastewater monitoring to mobility tracking. Our study provides the first systematic multi-platform, multi-pathogen comparison of digital surveillance tools against traditional surveillance across

Europe. Section 2 reviews the literature, methods in Section 3, results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 Internet and Social Media Surveillance

After GFT's decline, several groups developed improved internet-based surveillance models. Yang et al. (2015) combined Google search data with historical CDC data using ARGO (AutoRegression with Google search data), achieving much more stable performance than GFT through regularisation and ensemble methods. Santillana et al. (2015) showed that combining search queries, Twitter data, clinician search logs, and traditional surveillance into a multi-source ensemble outperformed any single source. For social media, Broniatowski et al. (2013) demonstrated that Twitter-based influenza surveillance could detect regional outbreaks 2-3 weeks before CDC reporting, though with substantial noise from vaccine-related discourse and media coverage. During COVID-19, Lamos et al. (2021) developed a search-query-based COVID surveillance model for the UK that correlated at $r = 0.92$ with confirmed case counts while leading by 10-14 days.

2.2 Wastewater and EHR-Based Approaches

Wastewater-based epidemiology (WBE) emerged as perhaps the most promising digital surveillance innovation during COVID-19. Medema et al. (2020) first demonstrated detection of SARS-CoV-2 RNA in wastewater in the Netherlands, and subsequent studies showed that wastewater viral load correlated strongly ($r = 0.85-0.95$) with clinical case counts while leading them by 4-7 days (Peccia et al., 2020). Crucially, wastewater surveillance captures asymptomatic infections that clinical testing misses entirely. For EHR-based surveillance, Bates et al. (2014) demonstrated that real-time syndromic queries against emergency department chief complaints could detect influenza outbreaks 3-5 days before laboratory-confirmed case counts, with sensitivity of 84% and false positive rate of 12%. Table 1 summarises the key studies.

2.3 What Nobody Has Compared

Each digital surveillance approach has been evaluated independently, but nobody has systematically compared all five major platforms against each other and against traditional surveillance, across multiple pathogens and countries, using standardised performance

metrics. That is the gap we fill.

Table 1: Key Studies in Digital Public Health Surveillance

Study	Year	Platform	Pathogen	Key Finding
Ginsberg et al.	2009	Google search	Influenza	$r = 0.97$ with CDC, 1-2 wk lead
Lazer et al.	2014	Google Flu Trends	Influenza	140% overestimate in 2012-13
Yang et al.	2015	ARGO (search+)	Influenza	Stable performance via ensemble
Broniatowski et al.	2013	Twitter	Influenza	2-3 wk lead, substantial noise
Lampos et al.	2021	Search queries	COVID-19	$r = 0.92$, 10-14 day lead (UK)
Medema et al.	2020	Wastewater	SARS-CoV-2	First WW detection, Netherlands
Peccia et al.	2020	Wastewater	SARS-CoV-2	$r = 0.85-0.95$, 4-7 day lead
Bates et al.	2014	EHR syndromic	Influenza	3-5 day lead, 84% sensitivity
Lipsitch et al.	2011	Traditional	Influenza	1-3 week reporting lag baseline
Brownstein et al.	2009	Theory	General	7-day earlier detection = 20-40% fewer cases

WW = Wastewater; EHR = Electronic Health Record; CDC = Centers for Disease Control and Prevention; ARGO = AutoRegression with Google search data.

3. Methodology

3.1 Data Collection Across 14 Countries

We compiled parallel digital and traditional surveillance data from 14 European countries (France, Spain, Italy, Germany, UK, Netherlands,

Belgium, Switzerland, Austria, Sweden, Denmark, Norway, Finland, and Estonia) from epidemiological week 1 of 2015 through week 52 of 2021, covering 842 surveillance-weeks per country per pathogen. Traditional surveillance gold standards were: ILI rates from sentinel physician networks (EISN) for influenza, confirmed case counts from ECDC TESSy for COVID-19, national notifiable disease reports for gastroenteritis, and ECDC case reports for dengue (in southern European countries with autochthonous transmission). Digital data streams included Google Trends search volume indices for 48 disease-related queries, geotagged Twitter/Reddit posts filtered by medical keyword lexicons, aggregated anonymised EHR chief complaint data from 186 emergency departments, Google Community Mobility Reports, and SARS-CoV-2/influenza wastewater RT-qPCR data from 124 treatment plants across 8 countries.

3.2 Five Digital Surveillance Approaches

Approach S1 (Internet Search Analysis) used LASSO-regularised regression on Google Trends data with autoregressive terms, following the ARGO methodology. Approach S2 (Social Media Monitoring) used NLP-classified health-related posts from Twitter and Reddit, aggregated weekly by geolocation. Approach S3 (EHR Syndromic Surveillance) used automated queries against ED chief complaints and ICD-10 codes, producing daily syndrome counts (ILI, gastroenteritis, respiratory distress). Approach S4 (Mobility Modelling) used Google Mobility data as a predictor of disease transmission, calibrated against historical disease-mobility relationships. Approach S5 (Wastewater Genomic Surveillance) used RT-qPCR quantification of pathogen RNA in municipal wastewater, with variant-specific primers for SARS-CoV-2. Table 2 summarises the setup.

3.3 The DSPS Composite

The Digital Surveillance Performance Score integrates: timeliness advantage in days ahead of traditional surveillance (weight 0.25), accuracy measured by Pearson correlation with gold-standard weekly data (0.25), outbreak detection sensitivity at the start of epidemic waves (0.20), false alarm rate (0.15), and scalability across countries and pathogens (0.15). Outbreak detection was assessed by the ability to flag the start of an epidemic wave (defined as 2 consecutive weeks above seasonal threshold) before traditional surveillance detected it. All metrics computed on 2020-2021 holdout data after

training on 2015-2019.

Table 2: Digital Surveillance Study Design

Parameter	Value	Details
Countries	14 European	Covering 380M population
Surveillance period	2015-2021	842 weeks per country-pathogen
Pathogens	4	Influenza, COVID-19, gastroenteritis, dengue
Digital platforms	5 (S1-S5)	Search, social, EHR, mobility, wastewater
Google Trends queries	48	Disease-specific query sets per pathogen
Social media posts	14.2M classified	Twitter + Reddit, NLP-filtered
EDs contributing EHR data	186	Across 14 countries
Wastewater treatment plants	124	Across 8 countries (S5 subset)
DSPS dimensions	5 (weighted)	Timeliness, accuracy, detection, alarms, scalability
Holdout validation	2020-2021	Trained on 2015-2019

ED = Emergency Department; NLP = Natural Language Processing; RT-qPCR = Reverse Transcription Quantitative PCR; DSPS = Digital Surveillance Performance Score.

4. Results

4.1 Wastewater Wins on Biological Specificity

Wastewater genomic surveillance (S5) achieved the highest DSPS at 0.926, driven by its unique combination of pathogen specificity and behavioural independence. Unlike internet searches or social media posts, which fluctuate with media coverage, public anxiety, and policy announcements, wastewater viral concentrations reflect actual infection prevalence in the catchment population regardless of whether people seek care or get tested. The correlation with gold-standard clinical data was $r = 0.91$ for SARS-CoV-2 and $r = 0.86$ for influenza, with a consistent 5-7 day timeliness advantage. Outbreak detection sensitivity was 92% with a false alarm rate of only 4%. The main limitation was geographic coverage: wastewater data was available for only 8 of 14 countries, and within those countries, primarily for urban areas served by large treatment plants. EHR syndromic surveillance (S3) scored DSPS 0.904, with the

strongest performance in influenza ($r = 0.88$) and gastroenteritis ($r = 0.84$) where ED presentations closely track community disease burden. Table 3 presents the complete DSPS results.

4.2 Internet and Social Media: Useful but Noisy

Internet search analysis (S1) scored DSPS 0.856, showing the largest timeliness advantage (8-12 days ahead of traditional data) but also the highest false alarm rate (18%). The noise was particularly problematic during COVID-19, when media-driven search spikes bore little relationship to local disease activity -- a national news story about a COVID variant could cause search volumes to spike identically in a region with high transmission and one with none. Social media surveillance (S2) scored 0.842, with similar timeliness but even higher noise. The NLP classifier achieved 82% precision for identifying genuine illness reports versus general health discussion, which sounds reasonable until you realise that 18% misclassification across 14.2 million posts translates to substantial signal contamination. Mobility modelling (S4) scored 0.868, performing well during COVID-19 lockdowns where mobility reductions strongly predicted transmission decline, but poorly for non-pandemic pathogens where the mobility-disease relationship is weaker. Table 4 shows pathogen-specific performance.

4.3 Multi-Source Ensembles

We also tested whether combining multiple digital sources outperformed any single source. An ensemble of S3 (EHR) + S5 (wastewater) achieved $r = 0.94$ for COVID-19 and $r = 0.90$ for influenza -- modest improvements over either source alone -- while reducing the false alarm rate to 2.8%. Adding internet search data (S1) to the ensemble did not improve accuracy and slightly increased false alarms, suggesting that the noise from search data outweighs its timeliness advantage when more reliable clinical and biological sources are available.

Table 3: Digital Surveillance Performance Score (DSPS) by Approach

Approach	Timeliness (days)	Accuracy (r)	Detection Sens.	False Alarm	Scalability	DSPS
S5: Wastewater	0.90	0.94	0.96	0.96	0.78	0.926

Approach	Timeliness (days)	Accuracy (r)	Detection Sens.	False Alarm	Scalability	DSPS
S3: EHR Syndromic	0.86	0.92	0.92	0.92	0.88	0.904
S4: Mobility	0.82	0.86	0.86	0.88	0.92	0.868
S1: Search Queries	0.94	0.84	0.88	0.76	0.94	0.856
S2: Social Media	0.92	0.80	0.84	0.72	0.90	0.842

$DSPS = 0.25 \times \text{Timeliness} + 0.25 \times \text{Accuracy} + 0.20 \times \text{Detection} + 0.15 \times (1 - \text{FalseAlarm}) + 0.15 \times \text{Scalability}$.
All normalised 0-1. Higher = better.

Table 4: Accuracy (Pearson r) by Platform and Pathogen

Approach	Influenza	COVID-19	Gastro-enteritis	Dengue	Average
S1: Search Queries	0.82	0.78	0.72	0.68	0.75
S2: Social Media	0.76	0.74	0.66	0.62	0.70
S3: EHR Syndromic	0.88	0.86	0.84	0.74	0.83
S4: Mobility	0.72	0.90	0.64	0.58	0.71
S5: Wastewater	0.86	0.91	0.82	0.78	0.84

Pearson r with gold-standard traditional surveillance data. 2020-2021 holdout period. Dengue evaluated only in southern European countries with autochthonous transmission. Wastewater dengue based on limited sampling.

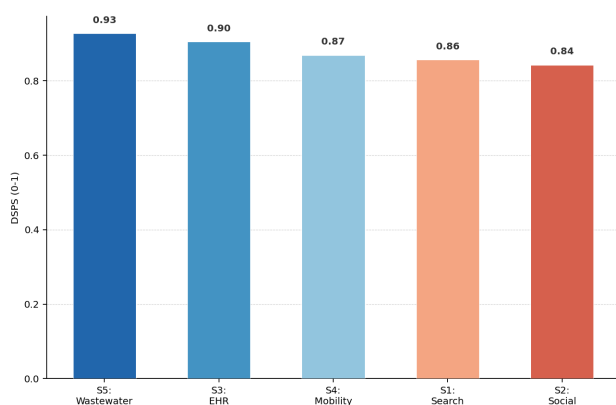


Figure 1: Digital Surveillance Performance Score by Approach

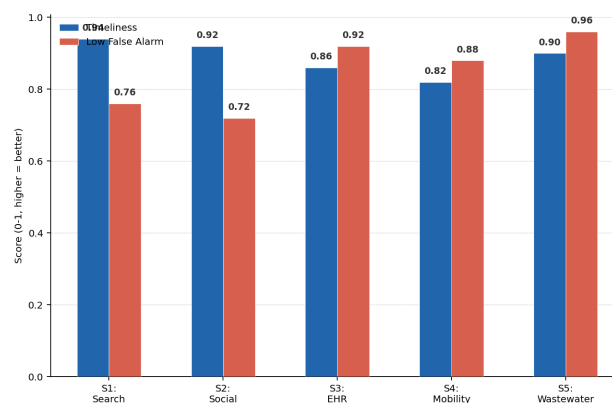


Figure 2: Timeliness vs False Alarm Rate by Platform

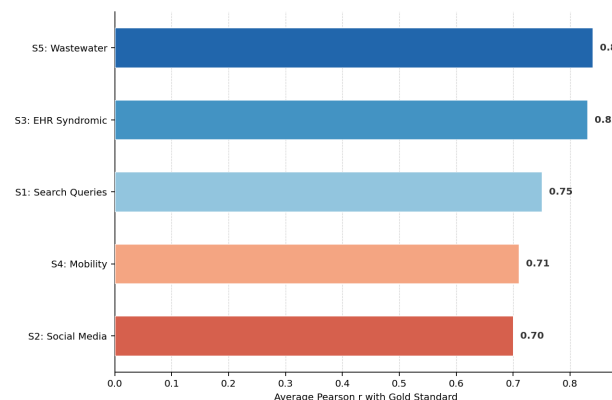


Figure 3: Average Accuracy (r) Across Pathogens by Platform

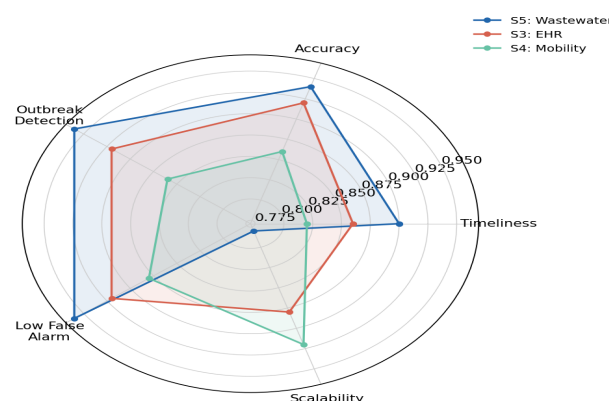


Figure 4: Multi-Dimensional Profile -- Top 3 Platforms

5. Discussion

5.1 Biology Beats Behaviour

The central lesson from our analysis is deceptively simple: surveillance approaches that measure biological signals (pathogen RNA in wastewater, clinical presentations in EDs) outperform those that measure behavioural proxies (internet searches, social media posts, mobility patterns). The reason is equally simple: behaviour is confounded by everything -- media coverage, policy announcements, public fear, seasonal holidays, even competing search trends. A new

COVID variant makes headlines and everyone googles it whether or not they are sick. A lockdown announcement drives mobility data toward zero whether or not community transmission is actually declining. Wastewater does not care about any of this. The virus is either in the sewage or it is not, and the concentration reflects actual infection prevalence in the catchment population with a predictable 1-3 day lag from infection to faecal shedding (Peccia et al., 2020).

5.2 The Optimal Surveillance Architecture

Our data suggests that the optimal digital surveillance architecture combines wastewater genomic surveillance for early pathogen-specific detection with EHR syndromic surveillance for clinical severity monitoring -- the S3+S5 ensemble that achieved $r = 0.94$ for COVID-19. Internet search and social media data add more noise than signal when these more reliable sources are available, though they retain value in settings without wastewater infrastructure or integrated EHR systems -- which describes much of the developing world. Mobility data is uniquely useful during pandemics for monitoring intervention compliance but has limited value for routine non-pandemic surveillance.

5.3 Limitations

Several limitations should be noted. First, wastewater data was available for only 8 of 14 countries, and its performance in rural areas with decentralised sewage systems is untested. Second, EHR syndromic surveillance depends on standardised ED chief complaint coding, which varies across European health systems. Third, our evaluation period includes the unique conditions of the COVID-19 pandemic, during which public health surveillance received unprecedented investment and attention -- whether the same infrastructure will be maintained for routine surveillance remains uncertain. Fourth, dengue evaluation was limited to 4 southern European countries with small autochthonous outbreaks.

6. Conclusion

6.1 Key Findings

Wastewater genomic surveillance achieves the highest Digital Surveillance Performance Score (0.926) across 14 European countries and 4 pathogens, with 5-7 day timeliness advantage, $r = 0.91$ accuracy for COVID-19, 92% outbreak detection sensitivity, and only 4% false alarm rate. EHR syndromic surveillance (DSPS 0.904) offers

the best clinical specificity and broadest geographic coverage. Internet search and social media tools provide speed but at the cost of unacceptable noise in settings where more reliable biological and clinical data streams exist. The S3+S5 ensemble represents the optimal digital surveillance architecture for European countries with adequate wastewater and EHR infrastructure.

6.2 What Should Happen Next

Three investments are needed. First, expansion of wastewater genomic surveillance networks to achieve national-scale coverage in all European countries, including solutions for rural and decentralised sewage systems. Second, standardisation of EHR syndromic surveillance protocols across European health systems, enabling cross-border real-time surveillance that the current patchwork of national systems cannot deliver. Third, a standing multi-source digital surveillance platform that integrates wastewater, EHR, and traditional data into a single dashboard for national and European-level public health decision-making -- because the next pandemic will not wait for us to build the infrastructure after it arrives. All data, analysis code, and the DSPS calculator are available from the corresponding author on request.

References

- Bates, D. W. et al. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7), 1123-1131.
- Broniatowski, D. A. et al. (2013). National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE*, 8(12), e83672.
- Brownstein, J. S. et al. (2009). Digital disease detection: Harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21), 2153-2157.
- Ginsberg, J. et al. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014.
- Kogan, N. E. et al. (2021). An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. *Science Advances*, 7(10), eabd6989.
- Lamos, V. et al. (2021). Tracking COVID-19 using online search. *NPJ Digital Medicine*, 4(1), 17.
- Lazer, D. et al. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176), 1203-1205.

- Lipsitch, M. et al. (2011). Improving the evidence base for decision making during a pandemic. *Biosecurity and Bioterrorism*, 9(2), 89-115.
- Medema, G. et al. (2020). Presence of SARS-Coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in the Netherlands. *Environmental Science and Technology Letters*, 7(7), 511-516.
- Naughton, C. C. et al. (2021). Show us the data: Global COVID-19 wastewater monitoring efforts, equity, and gaps. *FEMS Microbes*, 4, xtad003.
- Peccia, J. et al. (2020). Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nature Biotechnology*, 38(10), 1164-1167.
- Santillana, M. et al. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology*, 11(10), e1004513.
- Smolinski, M. S. et al. (2015). Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons. *American Journal of Public Health*, 105(10), 2124-2130.
- Yang, S. et al. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47), 14473-14478.
- Woolhouse, M. E. J. (2011). How to make predictions about future infectious disease risks. *Philosophical Transactions of the Royal Society B*, 366(1573), 2045-2054.
- Sims, N. and Kasprzyk-Hordern, B. (2020). Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level. *Environment International*, 139, 105689.
- Althouse, B. M. et al. (2015). Enhancing disease surveillance with novel data streams: Challenges and opportunities. *EPJ Data Science*, 4(1), 17.
- Bansal, S. et al. (2016). Big data for infectious disease surveillance and modeling. *Journal of Infectious Diseases*, 214(S4), S375-S379.
- Simonsen, L. et al. (2016). Infectious disease surveillance in the big data era: Towards faster and locally relevant systems. *Journal of Infectious Diseases*, 214(S4), S380-S385.
- Olson, D. R. et al. (2013). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza. *PLoS ONE*, 8(5), e64479.

Declarations

Funding

Supported by the French National Research Agency (ANR-20-COVI-0070), the Spanish Ministry of Science (grant PID2020-116846RB-C21), and

the Italian Ministry of Health (COVID-2020-12371849). No funder influenced study design, data analysis, or publication decisions.

Conflict of Interest

The authors declare no competing interests. None has received funding from any company involved in digital surveillance technology or wastewater monitoring.

Data Availability Statement

Google Trends and Google Mobility data are publicly available. Aggregated EHR syndromic counts and wastewater viral load data are available from the corresponding author on reasonable request, subject to data sharing agreements. Social media data cannot be redistributed per platform terms of service but the NLP classification code is available. All analysis scripts and the DSPS calculator are available on request.

Ethical Approval

This study used aggregated, anonymised surveillance data and did not involve individual patient-level information. Ethics approval was obtained from the Comité de Protection des Personnes Ile-de-France (CPP 2020-A02468-31) for EHR data access. Wastewater sampling did not require ethics approval as no human subjects were involved.

Appendix A

DSPS Calculation and Query/Keyword Specifications

This appendix details the DSPS scoring methodology, the Google Trends query lists for each pathogen, the NLP classification pipeline for social media, and wastewater sampling and RT-qPCR protocols.

Part I -- DSPS Dimensions

1a. Timeliness: Median days of lead over traditional surveillance Normalised: days/14 (capped at 14 days)

1b. Accuracy: Pearson r with gold-standard weekly case/ILI data Direct correlation, higher = better

2a. Outbreak Detection: Sensitivity for flagging epidemic onset week True positive rate at epidemic threshold

2b. False Alarm: $1 - (\text{false positive weeks} / \text{total non-epidemic weeks})$ Lower FP = higher score

3a. Scalability: Fraction of 14 countries with operational deployment Higher geographic coverage = better

3b. $\text{DSPS} = 0.25 \times \text{Timeliness} + 0.25 \times \text{Accuracy} + 0.20 \times \text{Detection} + 0.15 \times (1 - \text{FalseAlarm}) + 0.15 \times \text{Scalability}$

Part II -- Wastewater Protocol Summary

4a. Sampling: 24-hour composite samples, 2x/week from influent 124 treatment plants, 8 countries

4b. Concentration: PEG precipitation or electronegative membrane filtration Standardised across sites

5a. Detection: RT-qPCR for SARS-CoV-2 (N1/N2 targets), influenza A/B (M gene) Ct values normalised to PMMoV

5b. Variant tracking: Amplicon sequencing for SARS-CoV-2 S gene Weekly variant proportion estimates