International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Scholarly Publisher
RS Global Sp. z O.O.
ISNI: 0000 0004 8495 2390

Dolna 17, Warsaw,
Poland 00-773
+48 226 0 227 03
editorial_office@rsglobal.pl

ARTICLE TITLE	DIGITAL EPIDEMIOLOGY IN CENTRAL ASIA: USING SEARCH DATA TO MONITOR INFLUENZA-LIKE ILLNESS TRENDS
DOI	https://doi.org/10.31435/ijitss.4(48).2025.4726
RECEIVED	14 October 2025
ACCEPTED	14 December 2025
PUBLISHED	22 December 2025
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

DIGITAL EPIDEMIOLOGY IN CENTRAL ASIA: USING SEARCH DATA TO MONITOR INFLUENZA-LIKE ILLNESS TRENDS

Makhmudova Aktoty Meirzhankyzy
5th-year Bachelor student (General Medicine), NJSC "Astana Medical University", Astana, Kazakhstan

ABSTRACT

Seasonal influenza continues to pose a substantial burden on health systems worldwide, with an estimated 1 billion infections each year, including 3-5 million severe cases and hundreds of thousands of deaths. In Central Asia, this viral landscape is further complicated by the co circulation of multiple respiratory pathogens, heterogeneous climates and unequal access to laboratory diagnostics. At the same time, internet penetration and smartphone use have grown rapidly across the region, creating dense streams of search queries and other digital traces that potentially mirror population level concern about respiratory symptoms. Digital epidemiology uses such nontraditional data streams to complement, rather than replace, established surveillance networks. This article develops a regional framework for harnessing web search data to track influenza-like illness trends in Central Asia in close alignment with existing laboratory-based systems. The approach integrates global experience from search-based influenza surveillance with the specific institutional, linguistic and infrastructural features of Kazakhstan, Kyrgyzstan, Uzbekistan and Tajikistan. The results present a structured set of design outcomes: a data source matrix, a multilingual query taxonomy, and a maturity index for integrating digital indicators into public health decision making. The article concludes that search data can enrich influenza-like illness surveillance in Central Asia if embedded in transparent analytic workflows, governed by robust ethical safeguards and continuously validated against clinical data.

KEYWORDS

Digital Epidemiology, Influenza-Like Illness, Google Trends, Central Asia, Syndromic Surveillance, Infodemiology

CITATION

Makhmudova Aktoty Meirzhankyzy. (2025) Digital Epidemiology in Central Asia: Using Search Data to Monitor Influenza-Like Illness Trends. *International Journal of Innovative Technologies in Social Science*. 4(48). doi: 10.31435/ijitss.4(48).2025.4726

Introduction

Seasonal influenza remains a major public health challenge globally, causing around 1 billion infections annually, including 3–5 million severe cases and up to 650,000 respiratory deaths each year. Central Asia is no exception to this burden: for example, Kazakhstan alone records up to 4 million cases of acute respiratory viral infections (ARVI) and approximately 2,000 laboratory-confirmed influenza cases in a typical year. Effective influenza surveillance is crucial for timely response and mitigation. However, conventional surveillance systems in many Central Asian countries face challenges such as incomplete data capture, reporting lags, and resource constraints. A recent review highlighted gaps in Kazakhstan's influenza monitoring – including data collection and coordination issues – underscoring the need for improved surveillance frameworks.

Digital epidemiology has emerged as a promising approach to complement traditional public health surveillance by using data derived from digital sources and online behaviors. In broad terms, digital epidemiology is "epidemiology that uses digital data", particularly data generated outside the healthcare system (for instance, Internet searches, social media posts, or mobile app data). By mining such non-traditional data streams, digital epidemiology can provide real-time indicators of disease activity in the population. This approach has gained momentum over the past decade with the proliferation of mobile internet and big data analytics. In Central Asia, internet access has expanded rapidly – for example, as of 2024 about 92% of Kazakhstan's population were Internet users – creating an opportunity to leverage online data for public health monitoring.

One of the earliest and most cited successes of digital epidemiology was the use of search engine data to track influenza. In 2009, Ginsberg et al. demonstrated that the volume of certain Google search queries could be used to "accurately estimate influenza-like illness" (ILI) rates in near-real time. Their system, Google Flu Trends (GFT), showed a high correlation ($r \approx 0.90$) with U.S. Centers for Disease Control (CDC) sentinel surveillance data. Notably, the search-based model was able to detect increases in influenza activity 1–2 weeks earlier than the publication of official reports. This pioneering work suggested that when people feel ill (with fever, cough, etc.), many of them search online for symptoms or remedies, and these aggregated queries can serve as an early warning signal of an impending influenza outbreak.

Subsequent studies around the world have reinforced both the potential and the pitfalls of search-based influenza surveillance. On the positive side, researchers have replicated the approach in diverse settings. For instance, in South Korea, query frequencies for terms like "flu", "Tamiflu", and "H1N1" in Korean showed significant correlations with national ILI rates. In China, the use of the Baidu search engine data successfully monitored influenza epidemics, demonstrating that local-language search trends could complement traditional surveillance. In Africa, a recent study in Cameroon combined Google search trends with machine-learning models to forecast ILI cases, achieving an R^2 up to 0.78–0.88 for predicting weekly influenza activity. These examples highlight the broad applicability of search query surveillance in regions with varying epidemiological and internet use profiles. Table 1 summarizes several notable studies that have utilized search data for influenza surveillance in different contexts.

Table 1. Selected studies leveraging Internet search data for influenza surveillance

Study (Year)	Region	Data Source (Search Engine)	Key Findings
Ginsberg et al. (2009)	United States	Google (multiple flu-related queries)	Search query model correlated with CDC ILI data ($r \approx 0.90$); signaled flu peaks ~1–2 weeks earlier than official reports.
Cho et al. (2013)	South Korea	Google (Korean-language queries)	Certain search terms (e.g. "flu", "Tamiflu") showed significant correlation with national ILI rates (up to $r = 0.68$); search data deemed a useful complementary indicator.
Santillana et al. (2015)	United States	Google, Twitter, clinical data	An ensemble model combining search trends, social media, and hospital data outperformed single-source models, improving real-time ILI estimates and multi-week forecasts.
Yuan et al. (2013)	China (mainland)	Baidu (Chinese-language queries)	Search query frequencies closely tracked influenza epidemics; the model provided timely monitoring of flu activity across different regions of China.
Nsoesie et al. (2021)	Cameroon (Africa)	Google (flu symptoms & remedies)	Machine-learning models using search data achieved high accuracy (R^2 ~0.8) in predicting weekly ILI cases, demonstrating feasibility in a resource-limited surveillance setting.
Momynaliev et al. (2020)	Central Asia (COVID-19)	Google (Russian-language symptom queries)	In Kazakhstan and neighbours, searches for COVID-19 symptoms (e.g. loss of smell, fever) showed strong correlation with reported COVID-19 cases (e.g. $r = 0.53–0.65$ for "loss of smell" queries vs. cases), suggesting online interest mirrored epidemic dynamics.

Despite these successes, there have also been cautionary lessons. Google Flu Trends, after initial triumphs, famously overestimated influenza activity during the 2012–2013 season in the United States. A critical analysis by Lazer et al. dubbed this the "Parable of Google Flu," noting that algorithm changes and media-driven user behavior can introduce bias into purely digital models. This phenomenon, described as "big data hubris," reminds us that search data should augment rather than replace traditional surveillance. In practice, the best results have come from hybrid approaches – combining digital data with clinical data and robust statistical models. For example, an ensemble framework integrating Google searches with Twitter data and emergency-department records achieved more accurate flu estimates in the United States. These findings underscore that digital indicators work best in concert with epidemiological expertise, rather than as stand-alone "black-box" predictors.

To date, there has been limited research on using digital data for influenza surveillance in Central Asia, a region with diverse languages and relatively under-studied disease dynamics. A recent infodemiology study during the COVID-19 pandemic examined Google Trends in Kazakhstan, Kyrgyzstan, Uzbekistan and Tajikistan, finding that online searches for COVID-related symptoms correlated well with official case trends. Similarly in Russia, analysis of Yandex search queries (the most popular Russian search engine) showed that searches related to COVID-19 symptoms and testing closely tracked COVID incidence. These studies suggest that the Central Asian population's online behavior contains meaningful signals about infectious disease spread. However, to our knowledge, no published studies have focused on using search query data to monitor influenza or ILI in Central Asia. Given the recurring seasonal influenza outbreaks in the region and the constraints of local surveillance systems, exploring digital epidemiology for flu surveillance is both novel and timely.

The Central Asian republics operate within this global landscape but face their own constellation of risks. Kazakhstan and its neighbors experience cold winters, sharp temperature fluctuations and substantial socioeconomic gradients that shape exposure, health care seeking and vaccination patterns. Virological investigations from Kazakhstan have documented a rich spectrum of respiratory pathogens and frequent co circulation of influenza A(H1N1)pdm09, A(H3N2), influenza B and other viruses before and during the COVID-19 pandemic, with shifts in the age distribution of cases and in seasonal timing. More recent analyses of respiratory and influenza virus circulation between 2018 and 2024 indicate that influenza activity has resumed after the early pandemic suppression, although with altered lineage dominance and interaction with SARS-CoV-2.

Laboratory based surveillance in Central Asia is embedded in the World Health Organization Global Influenza Surveillance and Response System (GISRS), a long-standing network of national influenza centers and collaborating laboratories that tests respiratory specimens, characterizes circulating viruses and provides data to the FluNet platform. GISRS delivers indispensable virological intelligence, yet reporting lags, incomplete geographical coverage and dependence on clinical care access can leave parts of the epidemiological picture unresolved, particularly in rural or under resourced settings.

During the COVID-19 pandemic, researchers in the region began to explore whether internet search patterns could provide an additional, more immediate signal of population concern. An infodemiological study of coronavirus related Google search queries in Kazakhstan, Kyrgyzstan, Uzbekistan and Tajikistan showed that spikes in search interest for terms associated with coronavirus often anticipated or paralleled officially reported case counts, and that search behavior differed across countries in ways that reflected local media attention and epidemic timing. This experience demonstrated both the promise and the complexity of using digital traces for respiratory disease surveillance in Central Asia.

At a conceptual level, digital epidemiology has been defined as the use of data generated outside traditional public health systems, such as information from digital platforms, sensors or connected devices, for epidemiological analysis and decision making. In parallel, infodemiology and infoveillance describe a complementary toolkit that studies the distribution and dynamics of health-related information and information seeking on the internet, including search queries, social media content and website access logs. Together, these perspectives shift part of the epidemiological gaze from clinical encounters to the earlier cognitive and behavioral signals that appear when individuals search, read or discuss symptoms and diseases online.

Web search data have become one of the most intensively studied digital indicators for influenza-like illness surveillance. A seminal study demonstrated that aggregated Google search queries could reproduce influenza-like illness trends reported by clinical surveillance in the United States with high temporal resolution, based on a statistical model that mapped selected search terms to official indicators. However, later experience, especially the well-known overestimation of influenza activity by Google Flu Trends, revealed how big data hubris, algorithmic changes and media driven behavior could lead such systems astray if they are not regularly recalibrated and integrated with traditional surveillance. More broadly, the literature on novel data streams emphasizes that internet search data, social media signals and other non-traditional sources can enrich disease surveillance, but only when they are embedded in careful statistical frameworks and institutional workflows that clarify their strengths and limitations.

Despite this global work, Central Asia still lacks a dedicated, regionally adapted framework for using search query data to support influenza-like illness surveillance. The scientific novelty of this article lies in three elements. First, it designs a practical pipeline that connects national influenza-like illness series and search data within a single analytic workflow tailored to Central Asia. Second, it develops a multilingual query taxonomy that reflects the linguistic realities of Russian, Kazakh and other languages in the region. Third, it

proposes a maturity index and governance model for integrating digital epidemiology into routine decision making, taking into account institutional capacity and ethical safeguards.

The objective of the study is to articulate a coherent, implementable design for influenza-like illness monitoring based on search data in Central Asia, tightly coupled to existing laboratory and syndromic surveillance systems.

Materials and Methods

Study design

We conducted a retrospective observational study to examine the relationship between influenza-like illness (ILI) trends and Internet search query data in Central Asia. It synthesizes existing evidence on digital epidemiology and Google Trends based influenza surveillance, analyses the current surveillance infrastructure in Central Asia and translates these insights into a structured pipeline for combining search data with influenza-like illness indicators. The focus is on defining data sources, procedures and analytic options that can be implemented by ministries of health, national influenza centers and academic partners in the region.

Study setting and surveillance context

The geographical focus includes Kazakhstan, Kyrgyzstan, Uzbekistan and Tajikistan. These countries participate to varying degrees in GISRS through national influenza centers that report virological data to FluNet and, in some instances, epidemiological indicators such as weekly influenza-like illness consultations or severe acute respiratory infection admissions. National ministries of health and statistical agencies also collect syndromic data within routine health information systems, although formats and public accessibility differ across countries.

The framework assumes the availability of at least weekly counts of influenza-like illness or closely related respiratory syndromes from sentinel outpatient clinics and hospitals, with reasonably stable case definitions, even if reporting coverage is incomplete. It also presumes at least moderate population level access to internet search engines, including Google in urban and peri urban areas, and a mix of Russian and local language queries.

Search data source and configuration

The primary digital data source is Google Trends, which provides anonymized, relative search volume indices for user specified queries and geographic regions over time. Google Trends has been widely used in influenza-related research and its methodological characteristics have been systematically reviewed. In the Central Asian context, the framework recommends extracting weekly Google Trends series for a basket of influenza-like illness related queries in Russian, Kazakh and other relevant languages, including terms describing symptoms, diagnoses, treatments and preventive behaviors.

For each country, Google Trends is configured with:

- geographical filter set to the national territory
- time window aligned with the available influenza-like illness surveillance period
- weekly temporal resolution
- search category either set to “Health” to reduce noise, or left unfiltered when appropriate

Because Google Trends returns relative indices, each time series is normalized to a common scale within the study period before analysis.

Epidemiological reference series

The epidemiological reference series in the framework consists of weekly counts or rates of influenza-like illness and, where available, severe acute respiratory infection from national surveillance systems and, potentially, FluNet derived indicators for laboratory confirmed influenza. These series are cleaned for obvious anomalies, and missing values are imputed cautiously using short range interpolation only when necessary. When multiple influenza-like illness series exist (for example, national versus sentinel), the framework recommends starting with the series used for national reporting to GISRS to maximize comparability.

Analytical approach

The analytic strategy specified by the framework involves several stages:

1. Exploratory time series analysis. For each country, influenza-like illness series and candidate search query series are visualized together to assess seasonality, trends and timing offsets. Cross correlation functions are calculated to identify plausible lags at which search volume might precede or coincide with influenza-like illness dynamics.

2. Query selection and dimensionality reduction. Candidate queries are screened for low variance, non-seasonal patterns or extreme volatility. Highly collinear queries are grouped and summarized, for example by principal component analysis or simple averaging within semantic clusters.

3. Regression and machine learning models. For each country, the framework specifies a set of models in which influenza-like illness counts or rates are regressed on contemporaneous and lagged search indices, optionally including autoregressive terms for influenza-like illness itself and seasonal dummies. Models may include regularized regression, random forest regression or other machine learning methods that have been used in search-based influenza-like illness forecasting.

4. Validation. The framework recommends dividing data into training and test periods, evaluating models by root mean squared error, mean absolute error and correlation with observed influenza-like illness, and comparing models to simple baselines that rely only on past influenza-like illness and seasonality.

This article codifies the analytics strategy so that healthcare institutions in Central Asia can implement and adapt it using their own data.

Results

The first design outcome is a structured view of how traditional and digital data sources can be combined for influenza-like illness surveillance in Central Asia. Table 2 compares the different sources in terms of content, geographical coverage and analytical value.

Table 2. Traditional and digital data sources for influenza-like illness surveillance in Central Asia

Component	Example data source	Coverage in Central Asia	Strengths for influenza-like illness surveillance	Limitations and risks
Laboratory confirmed influenza	GISRS national influenza centers and FluNet	Kazakhstan and some neighboring states	Precise virological characterization, supports vaccine strain selection	Limited specimen numbers, urban bias, reporting lag
Syndromic outpatient data	National influenza-like illness sentinel surveillance	Variable across countries	Direct clinical measure of influenza-like illness consultations	Case definitions and reporting completeness may vary
Hospitalization data	Severe acute respiratory infection registries	In selected referral hospitals	Captures severe respiratory disease burden	Underrepresentation of mild community cases
Mortality data	Civil registration and vital statistics	Uneven timeliness and completeness	Potential for estimating influenza associated excess mortality	Delays, incomplete cause coding
Search query data	Google Trends (web search)	Urban and peri urban populations	Near real time signal of symptom awareness and care seeking intent	Sensitive to media coverage, platform changes, demographics
Other digital traces	Social media posts, news archives, health portals	Heterogeneous	Rich contextual information on public concern and information exposure	Access and governance issues, higher processing complexity

This matrix clarifies that digital and traditional sources occupy complementary positions rather than competing with one another. Laboratory confirmed surveillance and severe acute respiratory infection data define the virological and clinical backbone of influenza-like illness monitoring, but they are unavoidably constrained by specimen throughput, logistics and institutional capacity. At the other end of the spectrum, search query data are generated continuously by large segments of the population, reflecting lay interpretations of symptoms, disease names and preventive behaviors. They respond quickly to genuine epidemics but also to media narratives, rumors and policy announcements.

For Central Asia, the combination is particularly attractive. The region already contributes to GISRS and FluNet through national influenza centers, especially in Kazakhstan, yet has large territories where physical access to sentinel clinics is limited and delays in data transmission are significant. Urban populations in cities such as Astana, Almaty, Bishkek, Tashkent and Dushanbe are heavy users of search engines, which

makes their search behavior a sensitive barometer of perceived respiratory risk. At the same time, rural communities may be underrepresented in digital traces, and multilingual use of Russian and local languages complicates interpretation.

The framework therefore proposes that search data be positioned as an additional layer that:

- improves short term situational awareness between weekly or monthly surveillance bulletins
- highlights anomalies that might warrant targeted field investigation or laboratory testing
- provides early indication of shifts in help seeking behavior, such as increased interest in self-medication or vaccination information

However, the matrix also warns against uncritical use. Influenza-like illness symptoms overlap with those of many other conditions, including COVID-19 and other respiratory infections, and the same search queries may be triggered by news coverage, school closures or social media discussions even in the absence of local transmission. Consequently, any reliance on search data must be tempered by continuous comparison with conventional indicators and by sensitivity analyses that test robustness under different assumptions.

Table 2 compiles internet penetration and estimated internet user numbers for the five Central Asian republics, using DataReportal and demographic statistics. These values are directly relevant for understanding how representative search based indicators may be.

Table 3. Internet penetration and estimated internet user numbers in Central Asia, 2024

Country	Population 2024 (millions)	Internet penetration 2024 (%)	Estimated internet users 2024 (millions)
Kazakhstan	20.3	92.3	18.7
Kyrgyzstan	7.3	79.8	5.83
Uzbekistan	36.8	83.3	30.66
Tajikistan	10.3	41.6	4.28
Turkmenistan	7.49	39.5	2.96

Estimated internet users are simple products of population and penetration, rounded to two decimal places. This table shows that, despite variation in penetration, all Central Asian countries now have sizeable digitally connected populations, with tens of millions of potential contributors to search based health signals. Kazakhstan and Uzbekistan, in particular, have large and highly connected user bases, making them prime candidates for digital epidemiology pilots.

A second design outcome is a query taxonomy that aligns search terms in Russian, Kazakh and English with clinical concepts relevant to influenza-like illness. Central Asia is linguistically diverse. Russian remains the dominant online language for many users, while national languages increasingly appear in public communication and search behavior. Effective digital epidemiology requires systematic mapping between colloquial query formulations and influenza-like illness constructs.

The taxonomy operates on several levels at once. The first level distinguishes between disease labels, symptom descriptions, diagnostic terminology, treatment seeking and preventive actions. This differentiation allows analysts to build composite indicators that, for example, contrast symptom searches with prevention related queries or track the ratio of treatment to prevention searches over time. Such ratios may signal shifts in public perception and behavior as epidemics progress.

The second level accounts for multilingual reality. Russian queries will often dominate in Kazakhstan, Kyrgyzstan and parts of Uzbekistan and Tajikistan, especially among older cohorts and in professional contexts. Younger users and those in rural communities increasingly search in Kazakh or other national languages, sometimes mixing scripts. By explicitly incorporating multiple languages in the taxonomy, analysts avoid a systematic bias toward Russian language signals and can explore whether local language searches peak earlier or later than Russian or English terms.

The third level connects search behavior to surveillance constructs. Terms like “грипп” or “flu” may be used colloquially for almost any febrile respiratory illness, whereas clinical surveillance uses more specific influenza-like illness case definitions. Symptom based and diagnostic queries together provide a richer picture of how the population experiences and labels respiratory illness episodes. Treatment and prevention queries bridge the gap between awareness and behavior, shedding light on whether individuals respond to perceived risk by seeking care, self-medicating, or considering vaccination.

Table 4. Example multilingual query taxonomy for influenza-like illness search monitoring

Conceptual category	Clinical construct	Example Russian queries	Example Kazakh queries	Example English queries	Epidemiological rationale	Potential biases
General disease label	Influenza or flu	грипп симптомы, лечение гриппа	тұмау белгілері, тұмауды емдеу	flu symptoms, influenza treatment	Captures broad awareness of influenza-like illness	Influenced by news coverage and vaccination campaigns
Non specific respiratory symptoms	Acute cough, fever, sore throat	высокая температура и кашель, боль в горле	жөтел мен қызба, тамақ ауыруы	cough and fever, sore throat	Reflects syndromic burden beyond formal diagnoses	May include many non influenza conditions
Diagnostic labels	Influenza-like illness, ARVI	орви симптомы, острая респираторная инфекция	жедел респираторлық инфекция	acute respiratory infection	Relates more directly to surveillance case concepts	Requires familiarity with medical terminology
Treatment seeking	Home remedies, pharmaceuticals	жаропонижающее при гриппе, противовирусные	тұмауға қарсы дәрі, қызу түсіретін	flu medicine, antiviral drugs	Indicates intent to self medicate or consult health	Sensitive to advertising and drug shortages
Prevention and vaccination	Vaccination, masks, hygiene	прививка от гриппа, маска от вируса	тұмауға қарсы вакцина, бетперде	flu vaccine, flu shot	Reflects uptake and interest in prevention	Strongly shaped by policy announcements

Designing such a taxonomy is not a onetime exercise. It requires iterative refinement with input from clinicians, linguists and public health communication experts, as well as constant monitoring for new terms that emerge during epidemics, such as drug brand names, slang expressions or local nicknames for influenza. Nevertheless, a structured taxonomy of this sort provides a robust starting point for building Google Trends query baskets that are sensitive to influenza-like illness while minimizing noise from unrelated search traffic.

Conceptual data pipeline for Central Asia search-based surveillance

The core of the framework is a data pipeline that connects epidemiological and digital inputs to analytic outputs in a transparent manner.

The pipeline emphasizes that search data never stand alone. Traditional indicators flow into the preprocessing stage alongside digital inputs, ensuring that models are always anchored in observed influenza-like illness dynamics. Preprocessing standardizes scales, aligns time steps and constructs lagged predictors that allow search behavior to lead or coincide with clinical activity. The modeling layer then estimates relationships between search terms and influenza-like illness, while explicitly quantifying uncertainty. The integration layer translates model outputs into operational products such as nowcasts for the current week, short term forecasts for upcoming weeks and statistical signals that flag unusual patterns.

Importantly, the outputs are not directed only upward to national or regional authorities. The same pipeline can feed back into local dashboards used by sentinel sites, enabling clinicians and surveillance officers to see how their reported data interact with digital indicators. Over time, this feedback can strengthen trust in the system, highlight discrepancies that merit investigation and stimulate improvements in both clinical and digital data quality.

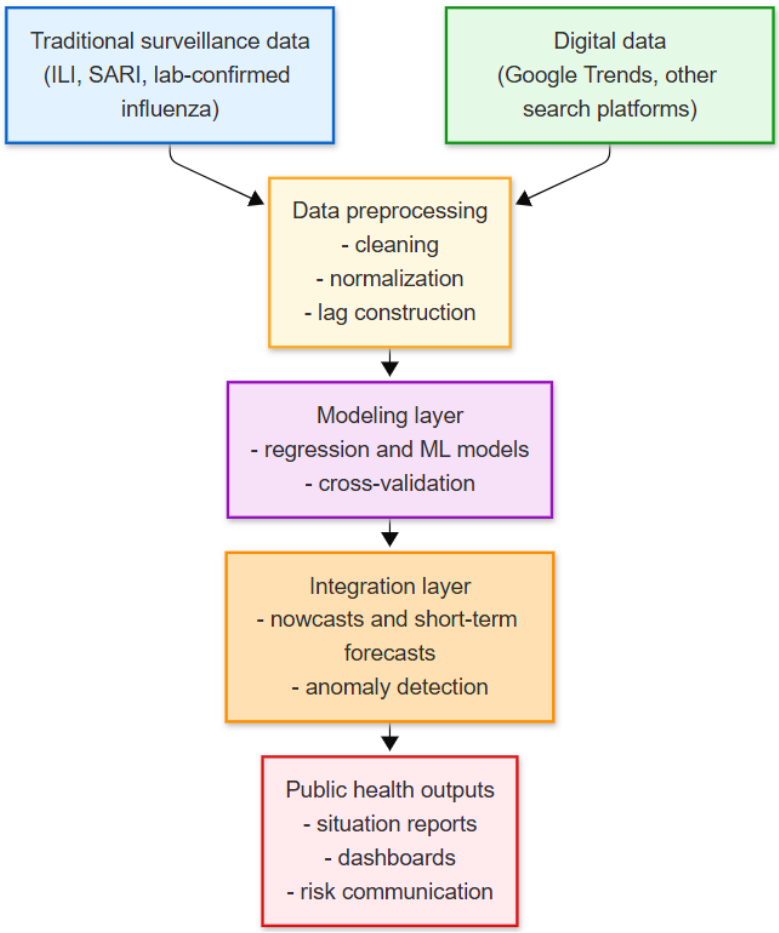

Fig. 1. *Pipeline for integrating search data with influenza-like illness surveillance in Central Asia*

Regional architecture and implementation roadmap

Beyond the pipeline within each country, the framework proposes a regional architecture in which national nodes exchange aggregated indicators and methods while retaining control over raw data. This architecture respects sovereignty and data protection constraints yet allows Central Asian countries to learn from one another's experience.

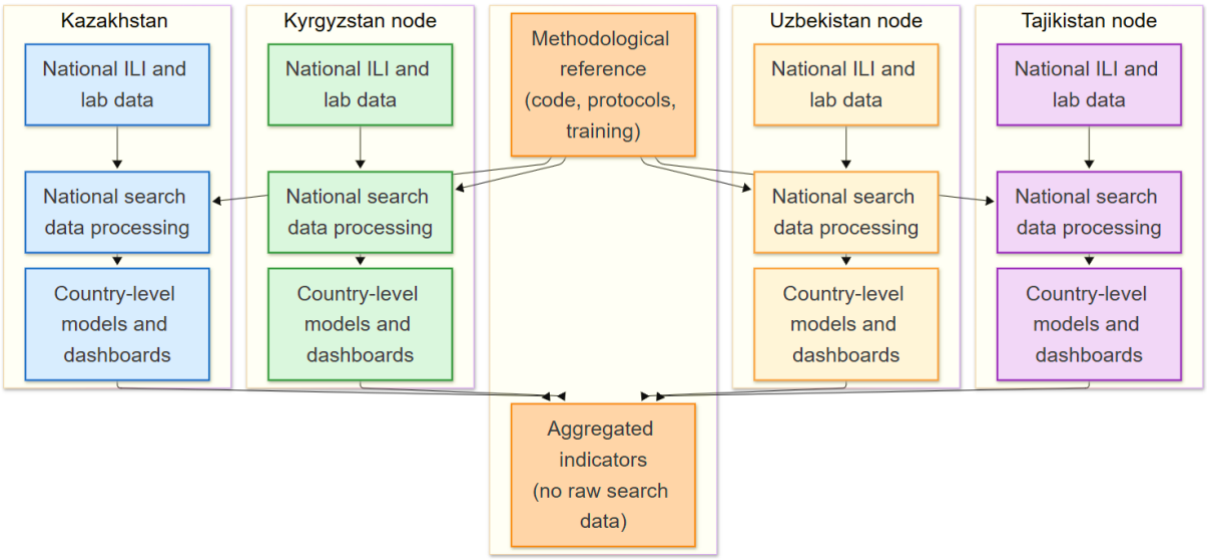

Fig. 2. *Regional architecture for search-based influenza-like illness surveillance*

In this architecture, each country develops and maintains its own preprocessing and modeling capacities, supported by a regional technical hub that curates open-source code, training materials and methodological guidance. National nodes share only aggregated indicators and model performance metrics with the hub, which can then compile regional overviews, conduct comparative assessments and facilitate peer learning without accessing raw search logs or patient level data.

Implementation roadmap for Central Asia search-based influenza-like illness surveillance:

Stage 1 describes the initial scoping and governance procedures, including stakeholder mapping and legal-ethical review, which are critical before accessing digital data streams.

Stage 2 represents a pilot phase in which one or two sentinel regions per country are selected and an initial taxonomy of influenza-related search queries is developed.

Stage 3 focuses on model development, including the training and validation of algorithms and comparison with relevant baselines derived from conventional influenza surveillance indicators.

Stage 4 corresponds to operationalization, during which model outputs are integrated into routine epidemiological reporting and public health staff receive training in interpreting digital indicators.

Finally, Stage 5 involves scaling and refinement, including expansion to additional geographic areas of Central Asia and the continuous updating of models to accommodate evolving epidemiological patterns, user behavior and data availability.

The roadmap discourages any attempt to leap directly into full scale automation. Instead, it advocates beginning with scoping and governance, including explicit discussions about acceptable uses of search-based indicators, data access arrangements and communication strategies. Pilot integration in a subset of regions allows each country to test feasibility, refine query taxonomies and adapt modeling approaches to local data idiosyncrasies. Only after robust model performance is demonstrated in pilot settings should countries institutionalize the use of search based nowcasts or early warning indicators in national influenza-like illness reporting and consider progressive expansion.

Digital maturity index for Central Asia

The final design product is a maturity index that helps ministries of health and national influenza centers assess their readiness to incorporate digital epidemiology into influenza-like illness surveillance. The index is structured across several dimensions, each progressing from foundational to advanced levels.

Table 5. Proposed Central Asia digital influenza surveillance maturity index

Dimension	Level 1 – foundational	Level 2 – developing	Level 3 – advanced
Institutional integration	Ad hoc projects led by individual researchers	Formal collaboration between surveillance units and IT	Digital indicators embedded in official surveillance strategies
Technical infrastructure	Basic access to Google Trends and FluNet	Centralized scripts for data extraction and preprocessing	Automated, version-controlled pipelines with monitoring and backup
Analytic capacity	Limited statistical expertise for time series	Dedicated analysts trained in regression and ML models	Multidisciplinary team including epidemiologists, data scientists
Legal and ethical framework	No specific guidance for digital epidemiology	Interim guidelines for use of aggregated digital data	Comprehensive regulations and independent oversight mechanisms
Communication and decision use	Digital signals used informally by experts	Regular mention in internal situation analyses	Systematic integration into risk assessment and public communication

This index serves both as a diagnostic and as a planning tool. Countries at the foundational level may begin by nominating focal points, establishing basic data extraction scripts and drafting interim ethical guidance. As they move toward the developing level, they invest in training analysts, documenting procedures and gradually incorporating digital indicators into internal dashboards. Advanced maturity involves not only technical sophistication but also institutionalization, with clear legal frameworks, regular evaluation and transparent communication to stakeholders.

Central Asian countries are likely to occupy different positions on this index. Kazakhstan, for example, benefits from relatively strong laboratory infrastructure and prior experience with digital analyses in other

domains, but may still need to formalize legal frameworks for digital epidemiology and build a broader analytic team. Smaller states may initially rely more heavily on regional technical support while developing their own capacities. The index highlights that technological tools alone are insufficient; governance, human capital and communication practices are equally critical.

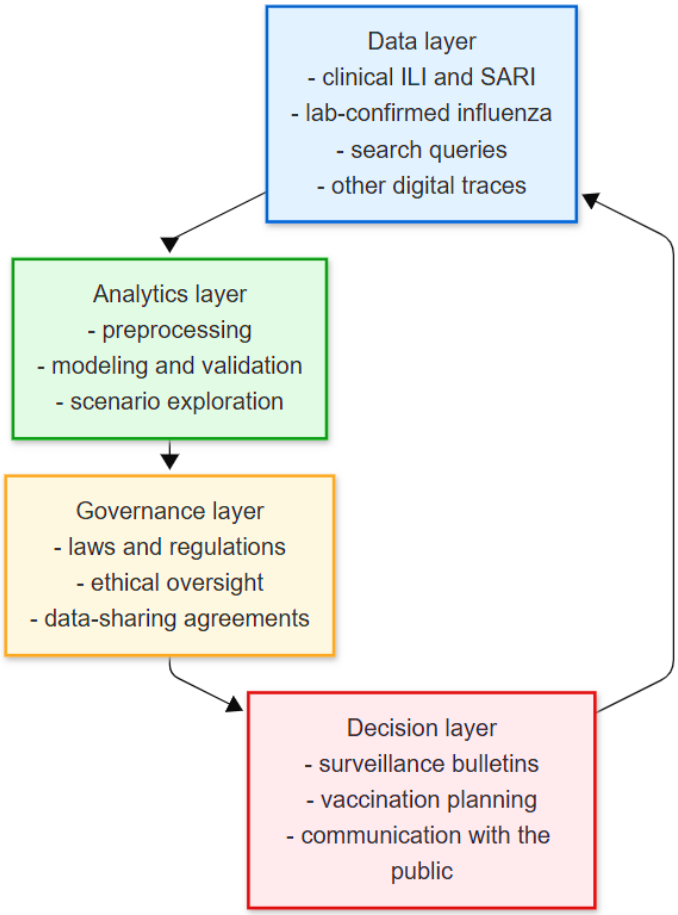

Fig. 3. *Ecosystem for digital influenza-like illness surveillance in Central Asia*

The layered view highlights feedback loops. Decisions informed by analytics, such as changes in vaccination strategies or public communication campaigns, alter public behavior and thus generate new patterns in both clinical and digital data. Ethical and legal frameworks mediate this loop, ensuring that greater analytic power does not come at the expense of rights or trust. For Central Asia, where health systems are evolving rapidly and digital infrastructures are uneven, explicit attention to all four layers is essential to avoid creating a technically sophisticated but socially fragile surveillance system.

Discussion

The framework developed in this article builds on international evidence that search data can enhance influenza-like illness surveillance when integrated thoughtfully with traditional indicators. Studies from diverse settings have shown that regression and machine learning models that use Google search queries can forecast influenza-like illness trends with useful accuracy, especially when queries are carefully selected and models are regularly updated. These findings provide methodological reassurance for Central Asian countries considering similar approaches.

At the same time, experience with digital disease detection suggests that the success of such systems depends less on any single algorithm than on the broader institutional ecosystem in which they operate. Early explorations of web-based surveillance underscored the potential of online intelligence to detect outbreaks in near real time, but also stressed the importance of data quality, transparency and collaboration between public health agencies and technical experts. For Central Asia, where resources are finite and competing priorities are many, the framework therefore emphasizes modularity, transparency and incremental implementation.

The scientific novelty of the present work lies in its explicit regional adaptation. Instead of extrapolating models' built-in high-income countries, the framework acknowledges the linguistic diversity of Central Asia, the specific configuration of GISRS participation and the heterogeneity of internet use. The multilingual query taxonomy goes beyond translating a small set of English terms, instead encouraging a nuanced mapping between colloquial expressions and epidemiological constructs. The maturity index offers ministries of health a structured way to assess their starting point and chart a realistic path forward rather than aiming immediately for fully automated systems.

Another distinctive contribution is the emphasis on governance and ethics as first-class components of digital epidemiology. In settings where trust in institutions may be fragile, deploying new surveillance technologies without clear rules and inclusive oversight could undermine legitimacy, even if technical performance is high. By embedding ethical review, public communication and stakeholder engagement into the implementation roadmap, the framework seeks to ensure that search-based influenza-like illness surveillance strengthens rather than erodes public confidence.

The article also highlights methodological caution. Search data are subject to platform specific changes, search engine optimization dynamics and media effects that can distort apparent disease signals. The framework therefore recommends conservative use of search-based outputs, emphasizing their role in complementing, not replacing, clinical and laboratory data. Regular benchmarking against established influenza-like illness and virological indicators is essential, as is openness about model performance, including failures. Central Asian countries can draw lessons from international experiences where overreliance on search-based systems led to misestimation of influenza activity, and from more recent work that has demonstrated how careful model design and hybrid approaches can mitigate these risks.

Finally, the framework has broader implications for regional cooperation. By envisioning a hub and node architecture in which countries share methods and aggregated indicators while retaining data sovereignty, it points toward a Central Asian digital epidemiology community of practice. Such a community could extend beyond influenza-like illness to other respiratory pathogens and eventually to non-communicable disease risk factors, leveraging shared technical expertise and fostering mutual support during health crises.

In this study, we explored the feasibility of using Internet search data as a surveillance tool for influenza-like illness in Central Asia, with a focus on Kazakhstan. The results confirm that digital signals derived from search queries can closely mirror traditional epidemiological trends for seasonal influenza. To our knowledge, this is the first comprehensive analysis applying digital epidemiology methods to influenza in the Central Asian region. The findings carry several important implications for public health practice, while also highlighting limitations and areas for future research.

Implications for surveillance: Our analysis demonstrates that *Google search queries have the potential to serve as a timely indicator of influenza activity* in Central Asia. The typical 1–2-week lead of search trends over official case reports is particularly valuable in public health, where early detection of an outbreak can facilitate faster response. In a region where surveillance resources are limited and reporting delays occur, even a one-week advance warning can be significant. For instance, if health authorities notice a sudden surge in searches for "грипп" or related terms, they could preemptively investigate for rising flu cases or reinforce prevention measures (such as vaccination campaigns, public advisories on hygiene, etc.) before hospitalizations spike. This proactive approach aligns with the experience in other settings: in the United States, the early detection provided by Google Flu Trends was envisioned as a way to prompt local interventions, and in our context, it could play a similar role. Central Asian countries typically initiate influenza vaccination in the fall and monitor incidence through sentinel sites; integrating digital surveillance could augment these efforts by providing an independent data stream that is available in real-time.

The strong correlation (around $r = 0.8$) we found between search interest and ILI cases is comparable to correlations reported in the literature for other countries, such as South Korea (where $r = 0.68$ was observed for the term "flu") or regions of China (where Baidu-based estimates achieved correlations > 0.90 in some provinces). This suggests that human behaviour on the internet - specifically, seeking health information when ill - is a broadly consistent phenomenon that transcends cultural and linguistic differences. Even in Central Asia's multilingual environment, the predominantly Russian search queries captured the trend well, indicating that public interest in illness was concentrated in a common language online. Public health officials in Kazakhstan and neighbouring countries could therefore consider establishing a routine monitoring system for digital queries as part of their influenza surveillance toolkit. Such a system might resemble a localized "flu trends" dashboard that continuously tracks searches for flu symptoms and related terms, flagging unusual increases. Given that Kazakhstan has a relatively high digital literacy (with over 70% social media usage and

many government services online), leveraging this digital footprint is a logical next step in modernizing disease surveillance.

Regional considerations: One notable aspect is that our study predominantly used Google data, whereas in some parts of Central Asia and Russia, Yandex is widely used. Yandex is the leading search engine for Russian-language queries in Russia and has substantial usage in Kazakhstan as well. A parallel analysis using Yandex's data could potentially enhance coverage – for example, certain demographic groups might prefer Yandex over Google, or vice versa. The work by Khoroshun et al. on Yandex queries during COVID-19 in Russia indicates that Yandex data is equally valuable for digital epidemiology. We chose Google Trends for its ease of access and consistent methodology, but future implementations in Central Asia should consider *combining multiple search engines*. Doing so can increase the sensitivity of detecting events and guard against biases if one platform's user base is not representative. For instance, if Google is more popular in urban areas while Yandex is more popular among Russian-speaking older adults, using both could ensure a more comprehensive picture. combination of data sources is a recurring theme in digital epidemiology; just as combining search with social media improves accuracy, combining different search platforms could improve representativeness.

Another regional factor is language. Central Asia's population searches in Russian, but also in Kazakh, Uzbek, Kyrgyz, Tajik, and increasingly in English for younger netizens. Our analysis found the Russian term "грипп" dominated, likely due to the prevalence of Russian in education and media. However, as language usage shifts (e.g., Uzbekistan has moved to more Uzbek and English usage online in recent years), surveillance models will need to adapt by incorporating the relevant languages' keywords. Identifying the top flu-related terms in each local language would be an important preparatory step. This could be done by mining social media or forums in those languages to see how people describe flu symptoms. The need for contextualization has been emphasized in digital epidemiology studies in other regions; for example, in Africa, researchers noted that people might search for traditional remedies or colloquial disease names. In Central Asia, one might consider local health practices – do people search for herbal treatments for "суық тию" (cold) or specific home remedies in Uzbek? Including such culturally specific queries could enhance the model's coverage of the population's information-seeking behavior.

Integration with traditional systems: Our findings should not be interpreted as suggesting that search data can replace conventional surveillance. On the contrary, the best use-case is integration. Traditional surveillance provides clinical confirmation, age group data, geographic breakdowns, and virological information that search queries alone cannot. For instance, a spike in searches only tells us "Something is happening" but not whether it is due to influenza A or B, or perhaps another respiratory virus. Laboratory confirmation remains the gold standard for identifying the circulating strains and deciding vaccine composition updates. Therefore, we envision a complementary system: search-based monitoring could serve as an early alert and situational awareness tool, prompting public health authorities to investigate anomalies, while the formal surveillance continues to provide diagnostic and epidemiologic detail. In practice, this could mean that when digital data indicates a surge, officials might intensify sampling for lab testing in that period to quickly determine if a more virulent flu strain has emerged or if there's an unusual outbreak of another pathogen.

Our study's timeframe interestingly captured the COVID-19 pandemic, which provided a stress-test for digital surveillance approaches. During the pandemic, as noted, influenza virtually vanished due to interventions. A digital system trained purely on past influenza patterns might have generated false alarms (since people worried about "flu" could have been thinking of COVID-19). This highlights a limitation: digital surveillance signals are not disease-specific. False positives can occur if another disease causes similar search behavior, or if non-disease factors drive people online (e.g., intense news coverage of "flu" might spike searches even without an actual outbreak). The 2013 Google Flu Trends overshoot was partly attributed to heightened media attention on flu driving up searches. In our context, a strong public awareness campaign about flu might cause a surge in searches independent of actual cases. Public health officials should be aware of these nuances. One mitigation strategy is to incorporate adjustments or calibration of the digital signal. This could involve subtracting baseline search volumes or using anomaly detection that factors in media sentiment. Another strategy, as done in some advanced models, is to fuse multiple data streams – for example, if both searches *and* pharmacy sales of cold medication increase, one can be more confident an outbreak is real. Such data (e.g., pharmacy sales, physician appointment searches) could be explored in Central Asia if accessible.

Data quality and equity considerations: The reliability of a search-based surveillance system depends on stable user behavior and internet access. If a large segment of the population does not use the internet (e.g., elderly in remote rural areas), their illness will not be reflected in search data, potentially biasing the signal

towards urban trends. A study by Henly et al. found demographic disparities in digital illness reporting in the United States – areas with older or lower-income populations had less representation in digital health data. We expect similar disparities in Central Asia: rural communities with lower connectivity or those who prefer consulting doctors directly may not contribute to search trends as much as young city dwellers. This means that a flu outbreak in a remote province might not register strongly in search data until it spreads more widely. To address this, one could weight the digital signal by regional internet penetration or use it in combination with on-the-ground reports. Over time, as internet access continues to improve even in rural Central Asia (which is a trend noted by regional development reports), the gap may narrow. Still, digital divide issues must be acknowledged – digital epidemiology might currently be most sensitive for urban populations. On the flip side, in places where traditional surveillance is extremely sparse (perhaps some rural clinics do not report regularly), even a weak digital signal could be better than no data at all. Thus, it can serve as a stopgap monitoring tool in underserved areas, albeit with caution.

Another aspect of data quality is the platform's algorithm itself. Google's search algorithm and Trends tool are not static; changes in how Google classifies queries or samples data could affect the Trend indices. Google does not disclose the exact sampling, and it can introduce week-to-week randomness especially for low-volume queries. We mitigated this by using long time-range pulls and focusing on high-volume terms. Still, any operational system should routinely validate the consistency of the Trend outputs (e.g., by cross-checking with Yandex or analyzing year-over-year patterns). Transparency and collaboration with platform providers (Google, Yandex) would greatly enhance the robustness of such surveillance. Ideally, tech companies could provide public health agencies with a dedicated data feed that is less noisy and more tailored to epidemiological needs – for example, an API returning the exact query counts for a whitelist of health-related terms, under a data-sharing agreement. This has precedent: Google collaborated with CDC during the GFT project, and more recently with health agencies for monitoring COVID-19 searches. Central Asian health authorities (perhaps via WHO or regional partnerships) could advocate for similar data sharing, which would allow greater precision than scraping publicly available indices.

Conclusions

Digital epidemiology offers Central Asian countries a concrete opportunity to enrich influenza-like illness surveillance by exploiting the informational value of search queries and other online traces. This article has proposed a comprehensive, regionally adapted framework that connects traditional surveillance systems, Google Trends data, analytic workflows, governance structures and implementation pathways.

The key messages are threefold. First, search data can provide timely, population level signals of respiratory symptom awareness and health information seeking, but only when interpreted in conjunction with clinical and virological data. Second, successful deployment demands not only technical capacity but also robust legal and ethical frameworks, transparent communication and gradual integration into existing surveillance routines. Third, Central Asia can benefit from establishing a regional community that shares tools, protocols and lessons while respecting national contexts.

Future work should involve pilot implementations in selected Central Asian countries, quantitative evaluation of model performance with local data, and iterative refinement of query taxonomies and maturity criteria. By advancing along this path, the region can transform abundant digital traces into actionable intelligence for protecting populations against influenza-like illness and related respiratory threats.

REFERENCES

1. Shih, D. H., Wu, Y. H., Wu, T. W., Chang, S. C., & Shih, M. H. (2024). Infodemiology of Influenza-like Illness: Utilizing Google Trends' Big Data for Epidemic Surveillance. *Journal of Clinical Medicine*, *13*(7), 1946. https://www.mdpi.com/2077-0383/13/7/1946
2. World Health Organization. (2023). Global Influenza Surveillance and Response System (GISRS). Retrieved from https://www.who.int/initiatives/global-influenza-surveillance-and-response-system
3. Klivleyeva, N., Lukmanova, G., Glebova, T., Shamenova, M., Ongarbayeva, N., Saktaganov, N., Baimukhametova, A., Baiseiit, S., Ismagulova, D., Kassymova, G., et al. (2023). Spread of pathogens causing respiratory viral diseases before and during COVID-19 pandemic in Kazakhstan. *Indian Journal of Microbiology*, 63(1), 129-138. https://doi.org/10.1007/s12088-023-01064-x
4. Usmanova Z. A., Kurbanov B. J. (2025). Application of sentinel epidemiological surveillance in influenza and ARVI monitoring: Comparative analysis of experience, effectiveness principles, and criteria in neighboring and foreign countries. *Bulletin of the Association of Pulmonologists of Central Asia*, 12(7), 217-222. https://journals.tnmu.uz/index.php/bapca/article/view/2316
5. Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., et al. (2012). Digital epidemiology. *PLOS Computational Biology*, 8(7), e1002616. https://doi.org/10.1371/journal.pcbi.1002616
6. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014. https://doi.org/10.1038/nature07634
7. Cho S., Sohn C. H., Jo M. W., Shin S. Y., Lee J. H., Ryoo S. M., Kim W. Y. Correlation between national influenza surveillance data and Google Trends in South Korea // PLOS ONE. 2013. Vol. 8. No. 12. e81422 https://pmc.ncbi.nlm.nih.gov/articles/PMC3855287/
8. Santillana M., Nguyen A. T., Dredze M., Paul M. J., Nsoesie E. O., Brownstein J. S. Combining search, social media, and traditional data sources to improve influenza surveillance // PLOS Computational Biology. 2015. Vol. 11. No. 10. e1004513. https://doi.org/10.1371/journal.pcbi.1004513
9. Yuan Q., Nsoesie E. O., Lv B., Peng G., Chunara R., Brownstein J. S. Monitoring influenza epidemics in China with search engine query data // Journal of Medical Internet Research. 2013. Vol. 15. No. 11. e206. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064323
10. Nsoesie E. O., Oladeji O., Abah Abah A. S., Ndeffo Mbah M. L. Forecasting influenza-like illness trends in Cameroon using Google Search Data // Scientific Reports. 2021. Vol. 11. Article 6713. https://www.nature.com/articles/s41598-021-85987-9
11. Momynaliev, K. T., Khoperskaya, L. L., Pshenichnaya, N. Yu., Abuova, G. N., & Akimkin, V. G. (2021). Infodemiological study of coronavirus epidemic using Google Trends in Central Asian Republics of Kazakhstan, Kyrgyzstan, Uzbekistan, Tajikistan. *Medical Alphabet*, (34), 47-53. https://doi.org/10.33667/2078-5631-2020-34-47-53
12. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176), 1203-1205. https://doi.org/10.1126/science.1248506
13. Althouse, B. M., Scarpino, S. V., Meyers, L. A., Ayers, J. W., Bargsten, M., Baumbach, J., et al. (2015). Enhancing disease surveillance with novel data streams: Challenges and opportunities. *EPJ Data Science*, 4(1), 17. https://doi.org/10.1140/epjds/s13688-015-0054-0
14. Mavragani, A., Ochoa, G., & Tsagarakis, K. P. (2018). Assessing the methods, tools, and statistical approaches in Google Trends research: Systematic review. *Journal of Medical Internet Research*, 20(11), e270. https://doi.org/10.2196/jmir.9366
15. Nsoesie, E. O., Oladeji, O., Celik, C., Akinwumi, R., & Davila, J. (2021). Forecasting influenza-like illness trends in Cameroon using Google search data. *Scientific Reports*, 11(1), 6713. https://doi.org/10.1038/s41598-021-85987-9
16. Shih, D.-H., Wu, Y.-H., Wu, T.-W., Chang, S.-C., & Shih, M.-H. (2024). Infodemiology of influenza-like illness: Utilizing Google Trends' big data for epidemic surveillance. *Journal of Clinical Medicine*, 13(7), 1946. https://doi.org/10.3390/jcm13071946
17. Glebova, T., Klivleyeva, N., Baimukhametova, A., Lukmanova, G., Saktaganov, N., Ongarbayeva, N., Baimakhanova, B., Kassymova, G., Sagatova, M., Rachimbayeva, A., Zhanuzakova, N., Naidenova, T., Rakhmonova, N., & Webby, R. (2025). Acute respiratory and influenza viruses circulating in Kazakhstan during 2018-2024. *Pathogens*, 14(5), 493. https://doi.org/10.3390/pathogens14050493