

Analytical workflow for outbreak origin and transmission assessment

1. Define objectives and hypotheses

Establish the key questions:

1. Where and when did illness-related activity first emerge?
2. How did it spread through the city over time?
3. Which transmission mechanism is most consistent with observed patterns?
 - Person-to-person/contact
 - Airborne or wind-influenced exposure
 - Waterborne exposure
 - Common-source event, such as a venue, food source, or transport hub
4. Where should public-health investigation and emergency resources be prioritized?

Treat microblog signals as an early-warning and hypothesis-generation source, not confirmation of infection.

2. Prepare and validate the data

Microblog data

- Remove duplicates, reposts, spam, bots, commercial content, and likely automated accounts.
- Standardize timestamps to one time zone.
- Validate GPS coordinates; remove impossible or low-accuracy locations.
- Aggregate messages into appropriate spatial units:
 - City zones for comparison with population data
 - A fine grid, such as 250–500 m cells, for hotspot detection
- Preserve original records for drill-down and audit.

Geographic and contextual data

- Load roads, highways, hospitals, landmarks, water bodies, and zone boundaries into a GIS.
- Calculate zone population and population density.
- Add weather records by time period, especially wind direction and speed.
- If available, add public-transit routes, schools, workplaces, event venues, wastewater infrastructure, and clinic/ER data.

Data-quality visual checks

Create maps and time-series plots for:

- Total posts by day and zone
- GPS point density
- Missingness or suspicious geographic concentration
- Posting behavior by account
- Weather changes over time

The analyst should inspect whether apparent hotspots are genuine signals or artifacts of high social-media use, tourist locations, hospitals, or large events.

3. Identify illness-related messages using NLP

Create a text-classification pipeline that separates relevant illness signals from unrelated discussion.

3.1 Develop a symptom/outbreak vocabulary

Include terms related to:

- Flu-like illness: fever, cough, chills, sore throat, body aches
- Severe illness: trouble breathing, pneumonia, hospitalization, death
- Household or social spread: “my family,” “coworkers,” “school,” “everyone at work”
- Potential exposures: tap water, restaurant, lake, event, bus, airport, hospital

Include slang, spelling variants, local language usage, negation, and sarcasm.

3.2 Classify message types

Assign each post to categories such as:

Category	Example interpretation
Self-reported illness	“I have fever and cough”
Illness in household/contact	“My child and husband are sick”
Observed local outbreak	“Everyone at school is sick”
News/repost/general discussion	“Flu cases rising in Vastopolis”

Category	Example interpretation
Death/severe outcome mention	“Neighbor died after being sick”
Exposure mention	“Got sick after the festival/water outage”
Irrelevant or ambiguous	Metaphorical or non-health use

Prioritize self-reports and direct reports of illness. Downweight news coverage because it can cause geographically widespread posting without reflecting local disease incidence.

3.3 Validate the classifier

- Manually label a representative sample of posts.
- Measure precision and recall, particularly for self-reported illness.
- Review false positives, such as “sick of work,” and false negatives caused by slang.
- Re-run classification after vocabulary refinement.

Visual inspection: show sampled posts on a map and in a timeline so analysts can verify that clusters reflect real reports rather than media attention.

4. Build a spatiotemporal illness-signal dataset

For each location and time interval, calculate:

- Number of probable illness posts
- Number of unique users reporting illness
- Illness-post rate per population
- Illness-post rate per active local users, if available
- Fraction of all posts that are illness-related
- Number of severe/death-related messages
- Symptom-specific rates, such as respiratory versus gastrointestinal symptoms

Use multiple time resolutions:

- Hourly, for early outbreak and event analysis
- Daily, for operational tracking
- Multi-day rolling averages, to reduce noise

Normalize by population and, where possible, baseline posting activity. Raw post counts alone can overstate activity in dense or highly connected areas.

5. Detect the likely outbreak origin

5.1 Temporal onset analysis

For every grid cell and city zone:

- Identify the earliest credible increase above baseline.
- Estimate the first sustained cluster rather than relying on a single isolated post.
- Plot epidemic curves by zone and compare timing.

Useful outputs:

- Animated daily heat map of symptom reports
- Map of estimated “first onset” time by grid cell
- Zone-level epidemic curves
- Timeline of severe illness and death mentions

5.2 Cluster and hotspot analysis

Apply spatial and space-time methods such as:

- Kernel density estimation for visual hotspot maps
- Space-time scan statistics to identify statistically unusual clusters
- Local Moran’s I or Getis-Ord G_i^* to detect high-high hotspots
- Change-point detection to identify sudden local increases

The probable origin is the earliest persistent, unusually high cluster after accounting for population and baseline posting behavior.

5.3 Human interpretation

Analysts should inspect the early cluster against:

- Hospitals, clinics, schools, transit hubs, workplaces, venues, and landmarks
- Water bodies, treatment facilities, and distribution networks
- Highways and major travel corridors
- Individual post text describing possible shared exposures

Do not designate an origin solely from a map hotspot. A hotspot near a hospital, for example, may reflect people posting while seeking care rather than the site of exposure.

6. Assess transmission mechanisms

Evaluate each mechanism against its expected spatial, temporal, and textual signature.

Hypothesis	Expected pattern	Key analyses
Person-to-person	Gradual spread outward from early clusters; household, school, workplace, and transit links; serial waves	Lagged correlations between adjacent zones; mobility/road-network spread; text references to contacts
Airborne/wind-influenced	Downwind excess cases after an apparent source; directional plume-like pattern; association with wind shifts	Overlay cases with wind vectors; upwind/downwind comparison; lagged wind regression
Waterborne	Cases aligned with water network, affected supply area, river/lake exposure, or shared source; often abrupt common onset	Compare rates by water-service area; inspect water-related text; assess gastrointestinal symptoms if present
Point-source/common exposure	Strong early concentration around an event, building, restaurant, market, or workplace; rapid onset among people with shared exposure	Landmark proximity analysis; event timelines; exposure keyword extraction; onset clustering
Healthcare-associated	Clusters around hospitals and healthcare facilities, possibly among visitors/ staff	Compare first-time self-reports versus hospital-related posts; inspect occupational and visit references

6.1 Person-to-person analysis

- Measure whether illness peaks move from one zone to neighboring or connected zones with plausible time lags.
- Build a network using adjacency, highways, transit corridors, or commuting flows.

- Test whether early hotspots predict later increases in connected locations.
- Search text for relationship and contact signals: family, office, classroom, roommate, bus, party.

A propagation pattern consistent with contact networks supports person-to-person spread.

6.2 Airborne/wind analysis

- For each time interval, map symptom-report density alongside wind direction and speed.
- Define downwind sectors from possible source locations.
- Compare illness rates downwind versus similarly populated upwind areas at plausible exposure lags.
- Check whether the spatial plume shifts when winds change.

A consistent downwind association across several days is stronger evidence than a single visual alignment.

6.3 Waterborne analysis

- Map reports relative to water bodies, water-treatment assets, and supply zones if available.
- Compare disease rates between populations served by different water systems.
- Extract references to tap water, drinking water, restaurants, swimming, flooding, or water outages.
- Examine symptom profiles: gastrointestinal symptoms may support waterborne exposure, though respiratory illness would weaken that hypothesis.

Avoid inferring waterborne spread from simple proximity to a river or lake; distribution infrastructure and exposure behavior matter more.

7. Integrate severe illness and death signals

Create a separate high-priority stream for posts mentioning:

- Hospitalization
- Severe respiratory symptoms
- Death
- Emergency services
- Vulnerable populations, such as elderly residents or care facilities

Map and time-order these signals relative to the general illness signal.

Interpretation:

- Severe reports concentrated after an earlier hotspot may indicate disease progression or delayed care.
- Severe reports concentrated around hospitals may reflect treatment location, not exposure location.
- Death-related posts should be manually reviewed because they may be news reports, rumors, or duplicated messages.

Immediately share credible high-severity clusters with public-health leadership for verification through hospitals, laboratories, and death surveillance systems.

8. Produce iterative decision products

Provide a regularly updated operational dashboard containing:

1. **Current illness intensity map** Population-adjusted illness-report rates and active hotspots.
 2. **Estimated outbreak progression map** First detection and direction of spread.
 3. **Transmission-hypothesis panel** Evidence for and against person-to-person, airborne, waterborne, and common-source explanations.
 4. **Severe-outcome map** Verified and unverified severe illness/death signals, clearly labeled.
 5. **Analyst review queue** High-impact posts, emerging clusters, unusual symptom combinations, and possible exposure locations.
 6. **Uncertainty indicators** Data coverage, classification confidence, low-volume zones, and potential media-driven reporting bias.
-

9. Recommend field investigation and resource deployment

Use the combined evidence to prioritize actions.

Finding	Recommended action
Early persistent cluster near a venue or workplace	Conduct exposure investigation, inspect site, identify attendees/workers

Finding	Recommended action
Spread follows social or transport connections	Expand testing, contact tracing, public messaging, and clinic capacity in linked zones
Downwind pattern is consistent across weather changes	Investigate environmental/industrial sources; issue exposure precautions while confirming
Cluster aligns with a water-service area	Test water supply, inspect treatment/distribution systems, issue boil-water or alternative-water guidance if warranted
Severe clusters exceed local healthcare capacity	Deploy mobile clinics, emergency transport, PPE, triage support, and surge staffing

10. Confirm findings with epidemiological data

Microblog analysis should guide rapid action, but the transmission mechanism and outbreak source require confirmation through:

- Clinical case reports and laboratory testing
- Hospital admissions and emergency-department syndromic surveillance
- Contact-tracing interviews
- Environmental sampling, such as water, air, or surface testing
- Pathogen sequencing, where available
- Wastewater surveillance
- Verification of reported deaths and their causes

The final conclusion should state the strongest supported hypothesis, competing explanations, evidence quality, and what data would confirm or refute each explanation.