Supplementary Methods

The Human Epidemic Database (HED) contains temporal and geographic case data on historic and ongoing infectious disease outbreaks. Historical data were originally collected for the 1918 and 1957 influenza pandemics and for events occurring during 1963 - 2016 with the following inclusion criteria:

1. the event was reported in the World Health Organization Disease Outbreak News (WHO DON) or
2. the event was caused by a viral pathogen and resulted in more than 50 deaths.

Other epidemiological studies have compiled temporal datasets from WHO DON reports [1], but our dataset has longer temporal coverage.

Near-real time surveillance of infectious disease outbreak events started in January 2017 to identify, assess, and collect data on epidemic events as they occur. Potential new events are identified daily through open-source, digital surveillance. Eligibility for inclusion in the HED is determined by human review of algorithmic scoring based on the following variables:

1. pathogen
2. geographic scale
3. epidemiological characteristics
4. total number of cases reported

For any event meeting the threshold for inclusion in the HED, our team identified the best available source(s) and structured data from all available reports dating back to the beginning of the event. After the initial structuring of an event-source in the HED, the digital surveillance team continues to monitor for newly published reports and adds new data to the HED as necessary until an event is declared over or 90 days elapses without any newly reported information.

To date (January 25, 2023), data have been collected from more than 500 distinct reporting sources comprising data for approximately 200 pathogens, 230 countries, and more than 3,150 distinct events; we analyze a subset of these events occurring through 2019 and excluding ongoing events . All data included in the HED are collected, structured, and validated as true-to-source. Data published by official (governmental or multilateral) reporting sources are given priority over other potential sources (such as media reporting). When necessary, sources published by international non-governmental organizations working in affected areas may also be collected. Traditional media sources are used only as a last resort, and social media sources are not used as data sources.

Data structuring follows methodologies to ensure consistency across all events in the HED. Structured data undergoes multiple rounds of peer review and automated validation to ensure data accuracy. The data structuring process is designed to produce the most reliable estimates possible of the distribution of reported cases and deaths over space and time. The data sourced from the HED that was used in this analysis is available on our github (https://github.com/concentricbyginkgo/zoonotic_spillover_trend/tree/master/data), along with the reporting source links used in the original data collection process for those events.

References

1. Torres Munguía, J.A., Badarau, F.C., Díaz Pavez, L.R. et al. A global dataset of pandemic- and epidemic-prone disease outbreaks. Sci Data 9, 683 (2022). https://doi.org/10.1038/s41597-022-01797-2