

1 **Title:**

2 **Epidemiological identification of a novel infectious disease in real time: Analysis of**  
3 **the atypical pneumonia outbreak in Wuhan, China, 2019-20**

4

5 **Sung-mok Jung<sup>a,1</sup>, Ryo Kinoshita<sup>a,1</sup>, Robin N. Thompson<sup>b,c,1</sup>, Katsuma Hayashi<sup>a</sup>,**  
6 **Natalie M. Linton<sup>a</sup>, Yichi Yang<sup>a</sup>, Andrei R. Akhmetzhanov<sup>a</sup>, Hiroshi Nishiura<sup>a,\*</sup>**

7 <sup>a</sup>Graduate School of Medicine, Hokkaido University, Kita 15 Jo Nishi 7 Chome, Kitaku,  
8 Sapporo, 0608638, Japan

9 <sup>b</sup>Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe  
10 Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK

11 <sup>c</sup>Christ Church, University of Oxford, St Aldates, Oxford OX1 1DP, UK

12

13 <sup>1</sup> Sung-mok Jung, Ryo Kinoshita, and Robin N. Thompson contributed equally.

14 \* Corresponding author at: Graduate School of Medicine, Hokkaido University, Kita 15  
15 Jo Nishi 7 Chome, Kita-ku, Sapporo-shi, Hokkaido 060-8638, Japan.

16 *E-mail address:* nishiurah@med.hokudai.ac.jp

17

18 **ABSTRACT**

19 **Objective:** Virological tests indicate that a novel coronavirus is the most likely  
20 explanation for the 2019-20 pneumonia outbreak in Wuhan, China. We demonstrate that  
21 non-virological descriptive characteristics could have determined that the outbreak is  
22 caused by a novel pathogen in advance of virological testing.

23 **Methods:** Characteristics of the ongoing outbreak were collected in real time from two  
24 medical social media sites. These were compared against characteristics of ten existing  
25 pathogens that can induce atypical pneumonia. The probability that the current outbreak  
26 is due to “Disease X” (i.e., previously unknown etiology) as opposed to one of the  
27 known pathogens was inferred, and this estimate was updated as the outbreak  
28 continued.

29 **Results:** The probability that Disease X is driving the outbreak was assessed as over  
30 32% on 31 December 2019, one week before virus identification. After some specific  
31 pathogens were ruled out by laboratory tests on 5 Jan 2020, the inferred probability of  
32 Disease X was over 59%.

33 **Conclusions:** We showed quantitatively that the emerging outbreak of atypical  
34 pneumonia cases is consistent with causation by a novel pathogen. The proposed  
35 approach, that uses only routinely-observed non-virological data, can aid ongoing risk  
36 assessments even before virological test results become available.

37 **Keywords:** Epidemic; Causation; Bayes' theorem; Diagnosis; Prediction; Statistical  
38 model

## 39 INTRODUCTION

40 A cluster of cases of atypical pneumonia with unknown etiology in Wuhan, China  
41 attracted global attention at the end of 2019 (Wuhan Municipal Health Commission,  
42 China, 2019; World Health Organization, 2020). An impressive series of rapid  
43 virological examinations ruled out common pneumonia-causing viruses such as  
44 influenza viruses, adenoviruses, and the coronaviruses associated with Middle East  
45 respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS) (Wuhan  
46 Municipal Health Commission, China, 2019; Normile, 2020a, 2020b; World Health  
47 Organization, 2020). As of 12 January 2020, the causative agent is suspected to be a  
48 coronavirus of non-human origin (European Centre for Disease Control and Prevention,  
49 2020; Normile, 2020b).

50 While examination of the viral genome is critical for identifying the pathogen,  
51 information made publicly available in real time describing clinical characteristics and  
52 other outbreak-related factors can also allow experts to consider the etiology and  
53 thereby differential diagnoses. For instance, most cases shared a history of visiting or  
54 working at a seafood market in Wuhan (Wuhan Municipal Health Commission, China,  
55 2020), where exposure to the novel coronavirus is suspected to have occurred with no  
56 evidence of direct human-to-human transmission (World Health Organization, 2020),  
57 leading us to believe that the cluster of cases was due to “Disease X” (i.e., an infectious  
58 disease of previously unknown viral etiology). However, rigorous quantitative  
59 assessment of the chance that the disease is in fact Disease X has not previously been  
60 undertaken. The present study addresses this, demonstrating that non-virological  
61 information can lead to an objective classification of Disease X, using a simple  
62 statistical model that exploits the well-known Bayes’ theorem.

## 63    **METHODS**

64    As the outbreak unfolded, we calculated in real-time the probability that the pathogen  
65    responsible for the atypical pneumonia was novel (Disease X), or whether instead the  
66    outbreak was generated by a previously known pathogen that can cause pneumonia. Our  
67    analysis began on 30 December 2019, when the Wuhan Municipal Health Commission  
68    announced that there had been a surprisingly large number of atypical pneumonia cases.  
69    At that time, we assumed the causative agent could have been one of seven known viral  
70    or three known bacterial diseases, along with the chance that it was instead Disease X.  
71    We tracked two of the most active medical social media sites, i.e., ProMED (ProMED,  
72    2020) and Flutracker (Flutracker, 2020), that reported the non-virological characteristics  
73    of the outbreak, including atypical pneumonia, other clinical characteristics, and  
74    exposure factors, as it progressed. These characteristics do not necessarily represent the  
75    features that were causing disease, but are instead basic observations from the ongoing  
76    outbreak. Given these characteristics, we then calculated the probability that the  
77    ongoing outbreak is due to a known disease or unknown Disease X. On the first day of  
78    calculation (i.e. 30 December 2019), the only explanatory factors we included was  
79    atypical pneumonia, which was common to all enumerated diseases. Our analysis  
80    represents simple logical deductions from the limited data that were available during the  
81    outbreak in a quantitative manner and was updated to reflect new information about the  
82    outbreak as it became available in real time.

83            Table 1 shows the information compiled about the current outbreak, and the  
84    dates on which each of these characteristics were discovered. Each characteristic listed  
85    was assigned a value of zero or one, denoting whether or not the characteristic of listed  
86    outbreak, not individual cases, was likely for the emerging outbreak, and the equivalent

87 values for outbreaks of previously observed pathogens were also noted. We make two  
88 assumptions to use and un-use a part of the input exposure characteristics: (i) previously  
89 known disease outbreaks are all based on empirically observed notion (and do not  
90 include the new exposure data (i.e., exposure at a wet market), that is specific to novel  
91 coronavirus in Wuhan, which may be non-informative to other outbreaks for the  
92 calculation) and (ii) all exposure characteristics are known for all previously known  
93 outbreaks, incorporating all factors enumerated. Also, once pathogens were ruled out as  
94 the causative agent of the current outbreak, they were removed from our analysis: for  
95 example, highly pathogenic avian influenza (HPAI) (H5N1) was confirmed not to be  
96 the causative agent by laboratory testing on 3 January. Hence, we omitted this pathogen  
97 from our analysis from 3 January 2020 onwards.

98 To assess the probability that the emerging outbreak was caused by a variant of  
99 a known pathogen, we first calculated the distance between the set of characteristics of  
100 the ongoing outbreak and those of previously known pathogens. The distance between  
101 the characteristics of the ongoing outbreak and cases due to pathogen  $j$  is denoted by  $d_j$ .  
102 We assumed that the probability that the outbreak is due to a variant of pathogen  $j$   
103 decreases exponentially with distance  $d_j$ . Then, by Bayes' theorem,

$$104 \quad Pr(\text{disease } j \mid \text{observed characteristics}) = \frac{Pr(\text{observed characteristics} \mid \text{disease } j)q_j}{\sum_i Pr(\text{observed characteristics} \mid \text{disease } i)q_i}, \quad (1)$$

105 in which the sum in the denominator is over all possible diseases  $i$  (i.e. each of the  
106 columns of Table 1, including the column describing the current outbreak). The  
107 constants  $q_i$  are *a priori* probabilities that the outbreak is due to pathogen  $i$  (Nishiura et  
108 al., 2012; Ejima et al., 2014). We set uninformative priors for all pathogens considered,  
109 so that  $q_i$  was simply the reciprocal of the number of pathogens being considered

110 (including Disease X) on each date in our analysis. We initially estimated the distance  
111 between observed characteristics of the outbreak and each known candidate pathogen  
112 using the Hamming distance (i.e., the sum of squares differences between the entries in  
113 the columns of Table 1 corresponding to the Disease X and the candidate pathogen).  
114 Then, we assumed that the probability that the outbreak is driven by disease  $j$  is  
115 governed by a negative exponential function

$$116 \quad \Pr(\text{observed characteristics} \mid \text{disease } j) \propto \exp(-d_j), \quad (2)$$

117 where  $d_j$  is the calculated Hamming distance.

118 We also repeated our analysis using an alternative measure of the distance  
119 between observed characteristics of the outbreak and each known candidate pathogen,  
120 namely the Euclidean distance (i.e. the square root of the Hamming distance). In each  
121 case, we assumed that the importance of each characteristic had an identical weight in  
122 our analysis, so that a simple quantitative assessment could be obtained in a  
123 probabilistic manner without the need for subjective judgement.

124 Combining equations (1) and (2), and assuming that  $q_i$  is identical over  $i$ , we  
125 have:

$$126 \quad \Pr(\text{disease } j \mid \text{observed characteristics}) = \frac{\exp(-d_j)}{\sum_i \exp(-d_i)}$$

127 The probability that the outbreak is driven by Disease X corresponds to the  
128 distance  $d_X = 0$ , and represents a risk score taking values between the reciprocal of the  
129 number of candidate pathogens including Disease X itself and one:

$$130 \quad \Pr(\text{Disease X} \mid \text{observed characteristics}) = \frac{1}{1 + \sum_{i \neq X} \exp(-d_i)}. \quad (3)$$

131           Supposing that there are  $n$  known pathogens responsible for the atypical  
132 pneumonia, the probability of observing Disease X without any information is identical  
133 with the probability of observing other listed pathogen (i.e.,  $1/(1+n)$ ) and as pathogens  
134 are ruled out by laboratory testing, the identical probability increases (i.e.,  $1/11$  until 2  
135 Jan 2020,  $1/7$  from 3 Jan 2020 and  $1/5$  from 5 Jan 2020 in current outbreak). In  
136 addition, if the probability of observing Disease X according to equation (3) takes a  
137 value close to the probability of observing other candidate pathogens, the overall  
138 probability that the outbreak is due to a novel pathogen should be interpreted as being  
139 low. A result of significant practical importance, however, is when the probability of  
140 observing Disease X is close to one or much larger than the probability corresponding  
141 to each previously observed candidate pathogen. In that case, all candidate pathogens  
142 are not similar to the causative agent of the ongoing outbreak, and so the outbreak is  
143 likely to be due to a novel pathogen.

144           We converted the probability of disease X into the equivalent percentage value  
145 (so that, for example, a result of 0.8 in equation (1) is assumed to mean an 80%  
146 probability) and refer to the percentage value as the “probability of Disease X”  
147 hereafter.

## 148 **RESULTS**

149 We show temporal changes in estimates of the probability that the ongoing outbreak is  
150 driven by each candidate pathogen in Figure 1. Because the only information on 30  
151 December 2019 was that cases displayed symptoms of pneumonia, the distance between  
152 ongoing outbreak and known ten diseases was all zero, and thus, all eleven candidate  
153 pathogens initially showed an identical probability of 9.1% (i.e.,  $1/11$ ). Additional

154 characteristics became known the following day (i.e., 31 December 2019), and  
155 consequently, the inferred probability that the outbreak was driven by a novel pathogen  
156 increased substantially to 58.6% and 36.9% for Hamming and Euclidean distance  
157 metrics, respectively. When the exposure characteristic (i.e. exposure at a wet market),  
158 that is specific to ongoing outbreak were excluded from the analyses, the probability of  
159 observing Disease X given observed characteristics is as high as 48.7% and 32.6% for  
160 Hamming and Euclidean distance.

161 Later in the outbreak, adenoviruses, HPAI (H5N1 and H7N9) and other  
162 influenza viruses were ruled out on 3 January 2020, leading the probability of Disease X  
163 being assessed as 90.7% and 57.2% for Hamming and Euclidean distance metrics, when  
164 all factors were considered as characteristic. Excluding the wet market exposure, the  
165 probability of Disease X was 78.2% and 50.6% for Hamming and Euclidean distance  
166 metrics, respectively. SARS- and MERS-associated coronaviruses were ruled out as the  
167 causative agent on 5 January 2020, leading to a very high estimate for the probability  
168 that the outbreak is caused by a novel pathogen once all information had been collected.  
169 As of 12 January 2020, the probability of Disease X is estimated to be 92.5% and 65.5%  
170 using the model considering all the factors, while the model excluding the characteristic  
171 of exposure at the wet market indicated that the probability of Disease X is assessed as  
172 81.8% and 59.1% for Hamming and Euclidean distance models, respectively

## 173 **DISCUSSION**

174 In this analysis, we have shown quantitatively that the ongoing outbreak of  
175 pneumonia cases in Wuhan has almost certainly been caused by a novel pathogen. This  
176 was demonstrated using a series of clinical, occupational, and behavioral observations



177 extracted from fragmented reports describing the cases as these reports became  
178 available in real time (European Centre for Disease Control and Prevention, 2020;  
179 Wuhan Municipal Health Commission, China, 2020). Although virological  
180 investigation is the gold standard for pathogen identification and a novel coronavirus  
181 has now been identified from some of the cases, such laboratory-based outcomes can  
182 only be obtained after successfully sequencing the novel virus, which can be a lengthy  
183 process. It still remains for the microbiological causal link to be established, for instance  
184 by ensuring that Koch’s postulates are met (e.g., as seen in a study of Zika virus (Krauer  
185 et al., 2017)). In the ongoing outbreak, the provisional identification of a novel  
186 coronavirus was performed on 7 January 2020 and announced formally on 9 January  
187 2020 (World Health Organization, 2020). We have shown that non-virological  
188 information can indicate that the cause of the outbreak is likely to be a novel pathogen,  
189 and that this conclusion could have been obtained before virological test results were  
190 announced. Disease X was inferred to be very likely on all dates from 31 December  
191 2019 onwards—the date on which descriptions of outbreak characteristics began to  
192 emerge.

193       When sufficient clinical details of cases (e.g., complete blood cell counts) are  
194 available, the number of causative pathogens considered can be limited to a reasonable  
195 number. In this instance, atypical pneumonia combined with reduced white blood cell  
196 counts and the lack of response to antibiotics indicated that the pathogen was consistent  
197 with viral rather than bacterial infection. With such information, collecting non-  
198 virological data can lead to a convenient quantification of the probability of Disease X,  
199 while awaiting the results of virological tests. We believe that the proposed approach

200 can greatly improve the ongoing risk assessment practices across the world.

201 It is critically important to discuss two issues that the definition of variables in  
202 Table 1 has involved. First, a critical underlying assumption is that Table 1 reasonably  
203 represents outbreak characteristics of ongoing and previously known outbreaks. The  
204 representation does not reflect observation from all confirmed cases nor epidemiological  
205 findings from a case control study (e.g. statistically significant risk factor). Rather, zeros  
206 and ones in the table were defined in a phenomenological manner. Depending on  
207 readers, the defined nominal values can be different from what it was shown in Table 1  
208 and ours is only for the exposition using a typical Table 1 that authors came up. Second,  
209 as we have shown, there are multiple combinations of characteristic data to be used.  
210 Namely, as an exposure to a wet market for known disease outbreaks other than HPAI  
211 was not necessarily derived from empirical observation, the fairness of an assumption  
212 that the majority of cases of those known disease outbreaks were asked not to have  
213 visited a wet market would be a subject for debate.

214 In the past, descriptive outbreak information has been used to produce sensitive  
215 outbreak case definitions, and causative agents have been pinpointed without using  
216 statistical methods in combination with epidemiological observations. In the present  
217 study, we have shown that such assessments can be made quantitatively using a simple  
218 statistical model, allowing for comparison of the likelihood of causative agents among  
219 all possible candidates. When outbreak characteristics are shared and updated in real-  
220 time (Table 1), these data can contribute to narrow down the possible range of causative  
221 agents. In the case of the outbreak in Wuhan, our calculation of the probability that each  
222 pathogen is the causative agent indicates that virologically excluding the possibility of

223 influenza viruses, adenoviruses and known virulent coronaviruses associated with  
224 SARS and MERS on 3 and 5 January 2020 can be regarded as an “unsurprising”  
225 finding.

226 As important limitations, the precision and credibility of input data, and the  
227 method for calculating the distance between candidate diseases and the observed  
228 outbreak, must be refined in the future. First, our proposed approach used very limited  
229 data in Table 1 for logical quantification of the probability that each pathogen was the  
230 causative agent. However, with more clinical data, the dataset of characteristics could  
231 be replaced by continuous frequencies (e.g. the frequencies of cases experience  
232 coughing and difficulty in breathing) rather than binary variables, and then the proposed  
233 method could even be used for screening suspected cases. Second, with such data it  
234 would also be possible to model the likelihood of a pathogen in equation (1) not by  
235 arbitrarily measuring the distance but by using classification models using regression or  
236 more sophisticated machine learning approaches. Third, the erroneous input of incorrect  
237 information may be a challenge in real time analyses, although this did not appear to be  
238 an issue during the course of our analysis of the outbreak in Wuhan. However, it must  
239 be considered that the veracity of the source of information for such an analysis could  
240 have an impact on the resulting probability calculations. Fourth, the estimated  
241 probability that an outbreak is driven by a novel pathogen might be slightly over- or  
242 underestimated due to limited information about the mode of transmission and small  
243 numbers of observed cases. Of note, we believe that without 100% specificity of  
244 bacterial pathogens linked to the ongoing outbreak, excluding bacterial pathogens as  
245 candidate cannot be ensured, while the chance that the current outbreak is due to

246 bacterial may be less suspected over time with partial clinical evidence. Nevertheless,  
247 the large number of characteristics that could be considered for the outbreak in Wuhan  
248 suggests that estimation was not beset in this study. Finally, we had to restrict ourselves  
249 to assume that the priori probability of all outbreak ( $q_i$ ) is identical. However, since the  
250 priori probability of observing the outbreak driven by a Disease X is completely  
251 unknown, we believe that this assumption can be plausible in this practice.

## 252 **CONCLUSIONS**

253 Despite the future improvements to our statistical modelling framework that  
254 are required, this short study has demonstrated clearly that the ongoing outbreak of  
255 pneumonia cases in Wuhan is consistent with causation by a novel pathogen, “Disease  
256 X.” Analyses of the type conducted in this study can greatly support virological and  
257 genetic efforts to characterize the causal agent of this and future outbreaks, with the  
258 benefit that such analyses can be carried out extremely quickly.

259

## 260 **Author’s contributions**

261 Sung-mok Jung: Data collection, formal analysis, model formulation, writing. Ryo  
262 Kinoshita: Data collection, formal analysis, visualization, writing. Robin N. Thompson:  
263 Data collection, model formulation, investigation, writing. Katsuma Hayashi: Data  
264 collection, visualization, writing. Natalie M. Linton: Data collection, model  
265 formulation, writing. Andrei R. Akhmetzhanov: Data collection, model formulation,  
266 writing. Yichi Yang: Data collection, writing. Hiroshi Nishiura: Conceptualization,

267 model formulation, supervision, fund raising, validation, writing.

## 268 **Acknowledgements**

269 The authors thank Dr Rebecca Spriggs for help devising the statistical approach. R.N.T.  
270 would like to thank Christ Church (Oxford) for funding via a Junior Research  
271 Fellowship. H.N. received funding from the Japan Agency for Medical Research and  
272 Development (AMED) [grant number: JP18fk0108050]; the Japan Society for the  
273 Promotion of Science (JSPS) KAKENHI [grant numbers, H.N.: 17H04701, 17H05808,  
274 18H04895 and 19H01074; R.K.: 18J21587], the Inamori Foundation, and the Japan  
275 Science and Technology Agency (JST) CREST program [grant number: JPMJCR1413].

276

## 277 **REFERENCES**

---

278 Ejima K, Aihara K, Nishiura H. Probabilistic differential diagnosis of Middle East  
279 respiratory syndrome (MERS) using the time from immigration to illness onset  
280 among imported cases. *Journal of Theoretical Biology*. 2014;346:47–53.  
281 European Centers for Disease Control and Prevention. Pneumonia cases possibly  
282 associated with a novel coronavirus in Wuhan, China.  
283 [https://www.ecdc.europa.eu/en/publications-data/pneumonia-cases-possibly-](https://www.ecdc.europa.eu/en/publications-data/pneumonia-cases-possibly-associated-novel-coronavirus-wuhan-china)  
284 [associated-novel-coronavirus-wuhan-china](https://www.ecdc.europa.eu/en/publications-data/pneumonia-cases-possibly-associated-novel-coronavirus-wuhan-china), 2020 (accessed 14 Jan 2020).  
285 Flutracker. [https://flutrackers.com/forum/forum/china-other-health-threats/china-](https://flutrackers.com/forum/forum/china-other-health-threats/china-emerging-diseases-other-health-threats/821830-china-41-diagnosed-viral-pneumonia-)  
286 [emerging-diseases-other-health-threats/821830-china-41-diagnosed-viral-pneumonia-](https://flutrackers.com/forum/forum/china-other-health-threats/china-emerging-diseases-other-health-threats/821830-china-41-diagnosed-viral-pneumonia-)

287 coronavirus-cases-in-wuhan-hubei-province-december-30-2019-1-fatal-case-739-  
288 screened-so-far-including-419-hcw, 2020, (accessed 14 Jan 2020).

289 ProMED. <https://promedmail.org/>, 2020, (accessed 14 Jan 2020).

290 Krauer F, Riesen M, Reveiz L, Oladapo OT, Martínez-Vega R, Porgo TV, et al. Zika  
291 Virus Infection as a Cause of Congenital Brain Abnormalities and Guillain–Barré  
292 Syndrome: Systematic Review. von Seidlein L, editor. PLoS Med.  
293 2017;14(1):e1002203.

294 Nishiura H, Mizumoto K, Ejima K, Zhong Y, Cowling B, Omori R. Incubation period  
295 as part of the case definition of severe respiratory illness caused by a novel  
296 coronavirus. Euro Surveill. 2012;17(42).

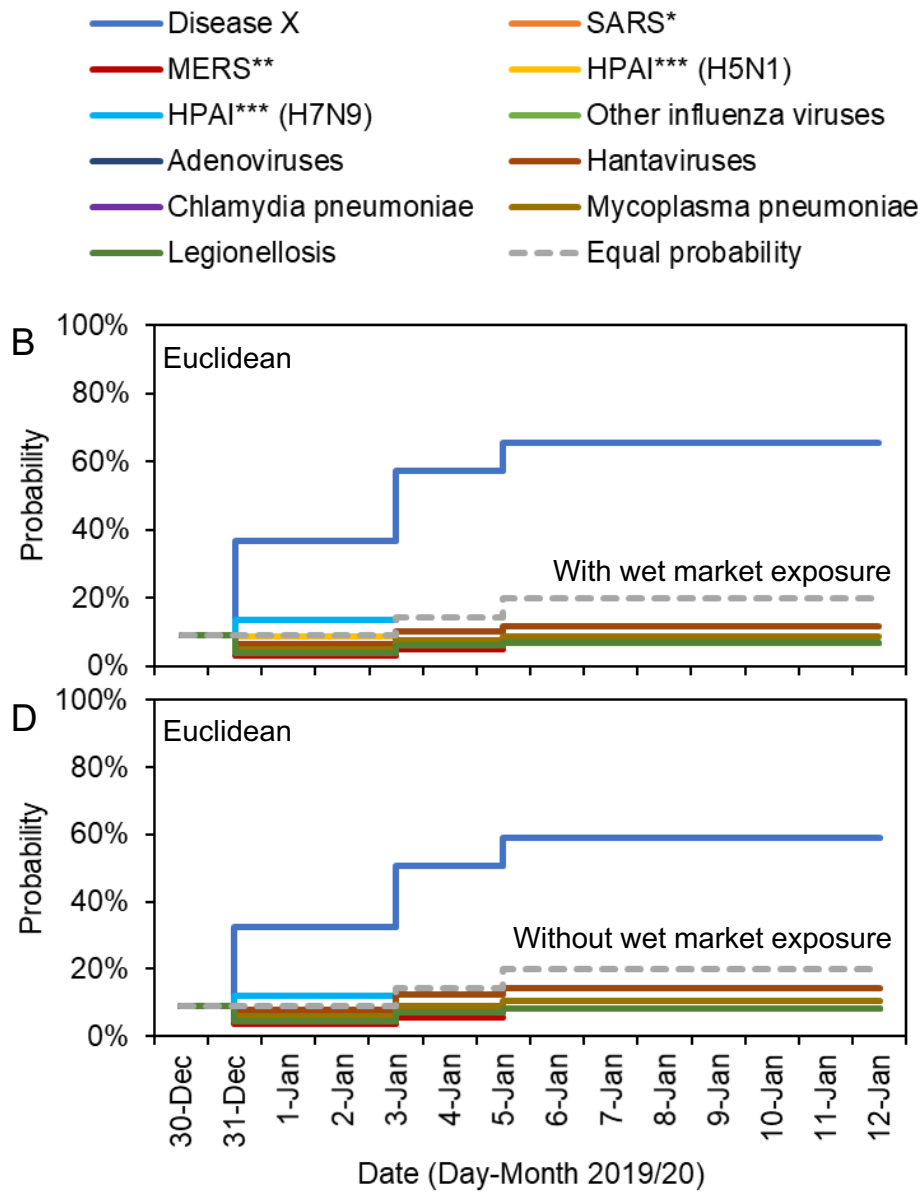
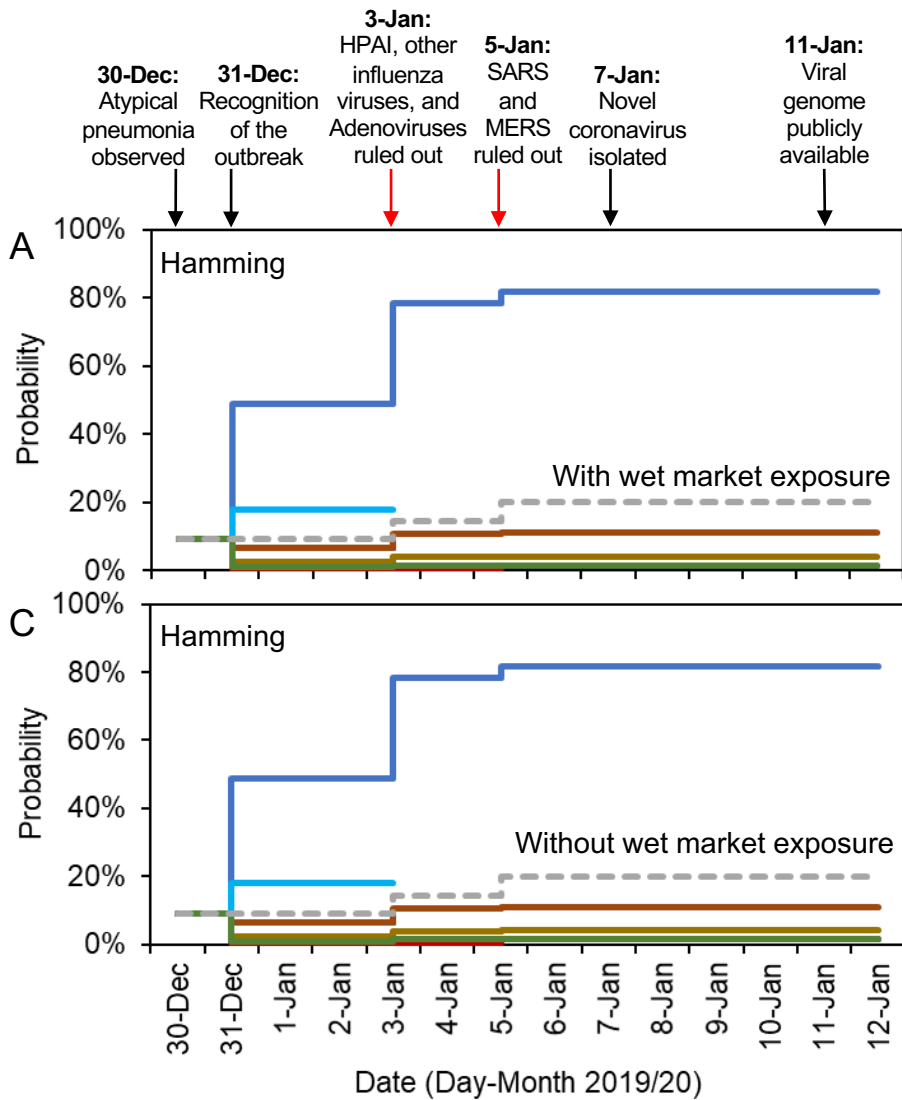
297 Normile D. Novel human virus? Pneumonia cases linked to seafood market in China stir  
298 concern. Science: New York. doi:10.1126/science.aba7672, available from:  
299 [https://www.sciencemag.org/news/2020/01/novel-human-virus-pneumonia-cases-](https://www.sciencemag.org/news/2020/01/novel-human-virus-pneumonia-cases-linked-seafood-market-china-stir-concern)  
300 [linked-seafood-market-china-stir-concern](https://www.sciencemag.org/news/2020/01/novel-human-virus-pneumonia-cases-linked-seafood-market-china-stir-concern), 2020a (accessed 14 Jan 2020).

301 Normile D. Mystery virus found in Wuhan resembles bat viruses but not SARS,  
302 Chinese scientist says. Science: New York. doi:10.1126/science.aba8542, available  
303 from: [https://www.sciencemag.org/news/2020/01/mystery-virus-found-wuhan-](https://www.sciencemag.org/news/2020/01/mystery-virus-found-wuhan-resembles-bat-viruses-not-sars-chinese-scientist-says)  
304 [resembles-bat-viruses-not-sars-chinese-scientist-says](https://www.sciencemag.org/news/2020/01/mystery-virus-found-wuhan-resembles-bat-viruses-not-sars-chinese-scientist-says), 2020b (accessed 14 Jan 2020).

305 World Health Organization. WHO Statement Regarding Cluster of Pneumonia Cases in  
306 Wuhan, China. World Health Organization: China.  
307 [https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-](https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china)  
308 [of-pneumonia-cases-in-wuhan-china](https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china), 2020 (accessed 14 Jan 2020).

309 Wuhan Municipal Health Commission. Wuhan Municipal Health Commission's  
310 briefing on the current pneumonia epidemic.  
311 <http://wjw.wuhan.gov.cn/front/web/showDetail/2019123108989>, 2019 (accessed 14  
312 Jan 2020).

313 Wuhan Municipal Health Commission, Wuhan Municipal Health and Health  
314 Committee's Report on Unexplained Viral Pneumonia.  
315 <http://wjw.wuhan.gov.cn/front/web/showDetail/2020010309017>, 2020 (accessed 14  
316 Jan 2020)  
317





318 **Figure legend**

319 **Figure 1. Real-time estimation of the probability that the ongoing pneumonia**  
320 **outbreak is driven by each candidate pathogen, given available information at**  
321 **different timepoints.** The probability that the outbreak is due to an unknown pathogen  
322 (Disease X) increases as more information becomes available, since the unknown  
323 pathogen can be seen to exhibit characteristics dissimilar to those observed in previous  
324 outbreaks, and since known pathogens are ruled out by laboratory results. Arrows  
325 indicate new information available on each date. Results are shown for different metrics  
326 describing the distance between characteristics of the ongoing outbreak and each  
327 candidate pathogen, and by knowledge (inclusion or exclusion) of exposure  
328 characteristics of Disease X (i.e. Work/visited a wet market), specifically: A. Hamming  
329 distance (the sum of squares difference between the entries in the columns of Table 1  
330 corresponding to the ongoing outbreak and the candidate pathogen considered) with wet  
331 market exposure; B. Euclidean distance (the square root of the Hamming distance) with  
332 wet market exposure; C Hamming distance without wet market exposure; D Euclidean  
333 distance without wet market exposure. Dashed grey line shows the probability without  
334 considering any information except atypical pneumonia (i.e. equal  
335 probability= $1/(1+\text{number of candidate pathogens})$ ). Note that the probability of some  
336 diseases is identical, for example, SARS and *Mycoplasma pneumoniae* has equal  
337 probability from 30 Dec to 4 Jan, and Legionellosis and *Chlamydia pneumoniae* has  
338 equal probability from 30 Dec to 12 Jan (Details in Supplementary material 1).

339

340 **Tables**

341 **Table 1. Characteristics of outbreaks driven by pneumonia-causing pathogens, with respect to the current outbreak in Wuhan,**  
 342 **China.**

Category	Characteristic	Current outbreak		Viral outbreaks							Bacterial outbreaks		
		Disease X	Date info shared	SARS*	MERS**	HPAI*** (H5N1)	HPAI*** (H7N9)	Other influenza viruses	Adenoviruses	Hantaviruses	<i>Chlamydia pneumoniae</i>	<i>Mycoplasma pneumoniae</i>	Legionellosis
Clinical	Atypical pneumonia	1	30-Dec	1	1	1	1	1	1	1	1	1	1
Clinical	CT (pulmonary infiltrates)	1	31-Dec	1	1	1	1	0	0	1	1	1	1
Clinical	Low white blood cell counts	1	31-Dec	1	1	1	1	1	1	1	0	0	0
Clinical	No response to antibiotics	1	31-Dec	1	1	1	1	1	1	1	0	0	0
Clinical	Frequent human transmission	0	31-Dec	1	1	0	0	1	1	0	1	1	1
Clinical	Substantial lethal cases	0	31-Dec	1	1	1	1	0	0	1	0	0	0
Travel/Occupation	Worked/visited a wet market	1	31-Dec	0	0	1	1	0	0	0	0	0	0
Travel/Occupation	Worked/visited a hospital	0	31-Dec	1	1	0	0	0	0	0	0	0	0
Travel/Occupation	Visited Middle East countries	0	31-Dec	0	1	0	0	0	0	0	0	0	0
Travel/Occupation	Visited hot spring or contact with potable water	0	31-Dec	0	0	0	0	0	0	0	0	0	1
Zoonotic	Contact with camels	0	31-Dec	0	1	0	0	0	0	0	0	0	0
Zoonotic	Contact with parrots/wild birds	0	31-Dec	0	0	1	0	0	0	0	1	0	0
Zoonotic	Contact with rodents	0	31-Dec	0	0	0	0	0	0	1	0	0	0

343 \*Severe acute respiratory syndrome; \*\*Middle East respiratory syndrome; \*\*\*Highly pathogenic avian influenza. Zeros represent characteristics that are unlikely  
 344 for outbreaks for that pathogen, and ones represent characteristics that occur. Dates and characteristics for the ongoing outbreak were obtained from two online  
 345 information systems [5,6], and information for other pathogens was summarised from the pathogen-specific pages on the WHO and CDC websites.