Ongoing diphtheria outbreak in Yemen: a cross-sectional and genomic epidemiology study

*Edgar Badell, Abdulilah Alharazi, Alexis Criscuolo, Khaled Abdullah Ali Almoayed, Noémie Lefrancq, Valerie Bouchez, Julien Guglielmini, Melanie Hennart, Annick Carmi-Leroy, Nora Zidane, Marine Pascal-Perrigault, Manon Lebreton, Helena Martini, Henrik Salje, Julie Toubiana, Fekri Dureab, Ghulam Dhabaan, Sylvain Brisse, NCPHL diphtheria outbreak working group**

Summary

Background An outbreak of diphtheria, declared in Yemen in October, 2017, is ongoing. We did a cross-sectional study to investigate the epidemiological, clinical, and microbiological features of the outbreak.

Methods Probable cases of diphtheria that were defined clinically and recorded through a weekly electronic diseases early warning system (from 2017, week 22, to 2020, week 17) were used to identify trends of the outbreak (we divided the epidemic into three time periods: May 29, 2017, to June 10, 2018; June 11, 2018, to June 3, 2019; and June 4, 2019, to April 26, 2020). We used the line list of diphtheria reports for governorate-level descriptions. Vaccination coverage was estimated using the 2017 and 2018 annual reports by the national Expanded Programme on Immunization. To confirm cases biologically, *Corynebacterium diphtheriae* was isolated and identified from throat swabs using standard microbiological culture and identification procedures. We assessed differences in the temporal and geographical distributions of cases, including between different age groups. For in-depth microbiological analysis, *tox* gene and species-specific *rpoB* real-time PCR, Illumina genomic sequencing, antimicrobial susceptibility analysis (disk diffusion, E-test), and the Elek diphtheria toxin production test were done on confirmed cases. We used genomic data for phylogenetic analyses and to estimate the nucleotide substitution rate.

Findings The Yemen diphtheria outbreak affected almost all governorates (provinces), with 5701 probable cases and 330 deaths recorded up to April 26, 2020. We collected clinical data for 888 probable cases with throat swab samples referred for biological confirmation, and genomic data for 42 positive cases, corresponding to 43 isolates (two isolates from one culture were included due to distinct colony morphologies). The median age of patients was 12 years (range 0·2–80). The proportion of cases in children aged 0–4 years was reduced during the second time period, after a vaccination campaign, compared with the first period (19% [95% CI 18–21] in the first period *vs* 14% [12–15] in the second period, p<0·0001). Among 43 tested isolates, 39 (91%) produced the diphtheria toxin and two had low level (0·25 mg/L) antimicrobial resistance to penicillin. We identified six *C diphtheriae* phylogenetic sublineages, four of which are genetically related to isolates from Saudi Arabia, Eritrea, and Somalia. Inter-sublineage genomic variations in genes associated with antimicrobial resistance, iron acquisition, and adhesion were observed. The predominant sublineage (30 [70%] of 43 isolates) was resistant to trimethoprim and was associated with unique genomic features, more frequent neck swelling (p=0·0029) and a younger age of patients (p=0·060) compared with the other sublineages. Its evolutionary rate was estimated at $1·67\times10^{-6}$ substitutions per site per year, placing its most recent common ancestor in 2015, and indicating silent circulation of *C diphtheriae* in Yemen before the outbreak was declared.

Interpretation In the Yemen outbreak, *C diphtheriae* shows high phylogenetic, genomic, and phenotypic variation. Laboratory capacity and real-time microbiological monitoring of diphtheria outbreaks need to be scaled up to inform case management and transmission control of diphtheria. Catch-up vaccination might have provided some protection to the targeted population (children aged 0–4 years).

Funding National Centre of the Public Health Laboratories (Yemen), Institut Pasteur, and the French Government Investissement d'Avenir Programme.

Introduction

Diphtheria is a severe infection that typically affects the upper respiratory tract, potentially leading to pseudomembrane formation, neck swelling, and suffocation.[1,2] In addition, diphtheria toxin production by some *Corynebacterium diphtheriae* strains can cause damage to the heart and other organs. Non-toxigenic strains can also cause invasive infections. Before large-scale vaccination, which mainly targets the toxin and is highly effective in preventing the disease, diphtheria was a major cause of death in children.[1] Large outbreaks of diphtheria are often observed after disruption to vaccination programmes, such as in ex-Soviet Union countries in the 1990s and more recently in Venezuela and the Rohingya refugee population in Bangladesh. Diphtheria is still observed occasionally in high-income countries, particularly among migrants.[3]

Lancet Microbe 2021; 2: e386–96

Published Online May 26, 2021 https://doi.org/10.1016/S2666-5247(21)00094-X

*Members of the NCPHL diphtheria outbreak working group are listed at the end of the Article

For the Arabic (Yemenite) translation of the abstract see Online for appendix 1

Biodiversity and Epidemiology of Bacterial Pathogens (E Badell PhD, V Bouchez PhD, M Hennart MSc, A Carmi-Leroy BTS, N Zidane PhD, M Pascal-Perrigault BTS, M Lebreton BTS, J Toubiana MD, S Brisse PhD), **National Reference Center for Corynebacteria of the Diphtheriae Complex** (E Badell, A Carmi-Leroy, J Toubiana, S Brisse), **and Hub de Bioinformatique et Biostatistique, Département Biologie Computationnelle** (A Criscuolo PhD, J Guglielmini PhD), **Institut Pasteur, Paris, France; National Centre of the Public Health Laboratories, Sana'a, Yemen** (A Alharazi MSc); **Department of Diseases Control and Surveillance, Ministry of Public Health and Population, Sana'a, Yemen** (K A A Almoayed MSc); **Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, UMR2000, Centre National de la Recherche Scientifique, Paris, France** (N Lefrancq MSc, H Salje PhD); **Collège Doctoral, Sorbonne Université, Paris, France** (M Hennart); **Department of Microbiology, National Reference Centre for Toxigenic Corynebacteria, Universitair Ziekenhuis Brussel, Vrije Universiteit Brussel, Brussels, Belgium** (H Martini MSc); **Department of Genetics, University of Cambridge, Cambridge, UK** (H Salje);

General Paediatrics and Paediatric Infectious Diseases Department, Université de Paris, Hôpital Necker-Enfants Malades, APHP, Paris, France (J Toubiana)**; Institute for Research in International Assistance, Akkon Hochschule, Berlin, Germany** (F Dureab PhD)**; Heidelberg Institute of Global Health, Heidelberg, Germany** (F Dureab)**; Mount Sinai Hospital, University Health Network, University of Toronto, Toronto, ON, Canada** (G Dhabaan PhD)

Correspondence to:
Dr Sylvain Brisse, Biodiversity and Epidemiology of Bacterial Pathogens, Institut Pasteur, Paris 75724, France
sylvain.brisse@pasteur.fr

or

Dr Ghulam Dhabaan, Mount Sinai Hospital, University Health Network, University of Toronto, Toronto, ON M5G 1X5, Canada
ghulam.dhabaan@utoronto.ca

Research in context

Evidence before this study

An outbreak of diphtheria has been ongoing in Yemen since October, 2017. We did a search of PubMed and Google on March 22, 2021, with no language restrictions, using the search terms "Yemen" and "diphtheria". This search identified that the epidemiology of the outbreak was described only up to mid-2018, and its microbiological and clinical features have not been reported. A search of PubMed for studies published from database inception up to March 22, 2021, using the term "diphtheria+outbreak+genomic", showed that only two outbreaks of toxigenic diphtheria have been investigated using genomic sequencing: an outbreak in South Africa in 2015, and a Belarus outbreak in 1996–2014. In both outbreaks, the coexistence of diverse sublineages of *Corynebacterium diphtheriae* and variable toxigenicity status were reported. Evolutionary rates were estimated for the Belarus outbreak sublineages ST5 and ST8: $5{\cdot}6\times10^{-7}$ substitutions per strain per year for ST5 and $8{\cdot}9\times10^{-7}$ substitutions per strain per year for ST8. Genomic sequences of *C diphtheriae* isolated in 2015 from refugees from Eritrea, Ethiopia, Syria, and Somalia were reported, but their relationship with Yemen isolates is unknown.

Added value of this study

To our knowledge, this is the first study to report epidemiological data collected since mid-2018 on diphtheria in Yemen, and to provide clinical features of cases of diphtheria from this ongoing outbreak. We defined three periods of the epidemic and showed that from the first period (May 29, 2017, to June 10, 2018) to the second period (June 11, 2018, to June 3, 2019), infection risk decreased in children aged 0–4 years. We report on the microbiological characteristics of the outbreak strains, including their toxigenic status, biovar, antimicrobial susceptibility, and virulence factor gene content, indicating variation in toxigenicity status and antimicrobial resistance, iron acquisition, and adhesin genes. These data indicate high heterogeneity of the strains within a single outbreak. We also estimate the evolutionary rate of the main Yemen *C diphtheriae* lineage, ST384 ($1{\cdot}67\times10^{-6}$ substitutions per site per year), and provide evidence for regional spread of *C diphtheriae* sublineages, based on genomic relatedness of *C diphtheriae* from Yemen with strains linked to Saudi Arabia and Eritrea.

Implications of all the available evidence

The combined epidemiological, clinical, and microbiological data from Yemen provide insight into one of the largest outbreaks of diphtheria in the past 20 years. Differences in the proportion of infections in children aged 0–4 years between the first and second period are consistent with some success from the mass immunisation catch-up programme. The evidence of genetic heterogeneity, antimicrobial susceptibility, and toxigenicity variation among *C diphtheriae* strains within a single large diphtheria outbreak shows the need for increased laboratory capacity and rapid testing, to improve diphtheria case management. Further investigation of both the local and regional spread of *C diphtheriae* will inform infection prevention strategies.

Genomic analysis of *C diphtheriae* isolates from outbreaks or surveillance programmes is critical in tracing transmission at local or regional scales, and in defining evolutionary rates, temporal depth of transmission, and genomic heterogeneity among strains.[3–5] Besides vaccination, a critical component of diphtheria management is serotherapy. Unfortunately, diphtheria antitoxin is not readily available in Yemen, as in most countries.[6] Antimicrobial treatment is also indicated for patient care and to avoid transmission, with penicillin and erythromycin being recommended as first-line therapeutic agents.[1,7]

In Yemen, where civil war has been ongoing since March, 2015, a large outbreak of diphtheria was recognised in October, 2017.[8–11] Although a vaccination campaign targeting 300 000 children began in late November, 2018, with plans to scale up to 3 million children and young adults in December, 2018, the conflict has complicated vaccination catch-up efforts.[12,13] Before this outbreak, diphtheria was considered endemic in Yemen, with an average of 50 suspected cases reported annually since 2000.[10] The last documented outbreak of diphtheria in this country occurred in 1981–82, with a total of 149 cases and no deaths. Yemen is currently also affected by other epidemic diseases, including cholera[14] and COVID-19.[15,16] As of May, 2020, the diphtheria outbreak was ongoing.[17] Although the early epidemiology of the outbreak was described,[8,9] almost no data have been reported on the clinical and microbiological characteristics of the outbreak. Here, we report on the clinical, epidemiological, and microbiological features of the ongoing Yemen diphtheria outbreak.

Methods

Study design

We did a cross-sectional study, including a molecular epidemiology investigation, of a diphtheria outbreak across Yemen, in 2017–20. For epidemiological investigations, several national-level data sources were used. Yemen is divided into 22 governorates, which are further divided into 333 districts. In this study, the majority of districts that provided epidemiological data were located within governorates in the central and northern regions of Yemen. In each district, a rapid response team of the surveillance department was mobilised through warning reports from health facilities. This team then both investigated the clinical cases and collected the samples for confirmation. Throat swabs were done for every tenth case if there was a cluster of more than ten cases. In

parallel, the electronic diseases early warning system (eDEWS) questionnaire was filled by the eDEWS focal person in each health facility on a weekly basis.[18] This sampling strategy remained stable over time.

Probable cases were defined by the rapid response team based on clinical examination showing an adherent membrane on one or more of the tonsils, pharynx, and nose, and any one of the following: laryngitis, pharyngitis, or tonsillitis.[18] We compiled probable cases from eDEWS releases from 2017, week 22 (May 29, 2017), to 2020, week 17 (April 26, 2020).[18] We used the line list of diphtheria reports for governorate-level descriptions. Vaccination coverage was estimated using the 2017 and 2018 annual reports by the national Expanded Programme on Immunization, obtained from the Ministry of Public Health and Population of Yemen.

Confirmed cases were defined as a probable case from which a *C diphtheriae* isolate was cultivated at the Yemen National Centre of the Public Health Laboratories (NCPHL).

To describe characteristics of the outbreak over time, we divided the epidemic into three time periods: May 29, 2017, to June 10, 2018; June 11, 2018, to June 3, 2019; and June 4, 2019, to April 26, 2020.

This work was done as part of routine case management under an emergency response mandate from the Government of Yemen and was supervised by the scientific committee of the NCPHL. The NCPHL collected samples and isolated strains and transferred some of them to the French National Reference Centre for Diphtheria for genomic sequencing.

Procedures

Throat swabs were stored in Amies transport medium (HiMedia, Mumbai, India) and transported within 24 h to the NCPHL laboratory on cold bags. Throat swab samples were analysed using standard microbiological procedures, as detailed in appendix 2 (pp 2–5). Briefly, swabs were plated on blood agar and tellurite blood agar (Sigma-Aldrich, Darmstadt, Germany). Gram and Albert stains were done to examine the presence of Gram-positive bacilli and metachromatic granules. Suspected *C diphtheriae* colonies were further sub-cultured on Tinsdale agar (Himedia, Mumbai, India), Dorset egg medium (Oxoid, Basingstoke, UK), and tellurite blood agar, followed by biochemical testing. Among stored positive cultures available in July, 2019, 100 samples were selected using a uniform random number generator. 98 of these stored samples were retrieved from storage (two tubes corresponding to selected samples had been lost) and were sent to Institut Pasteur (Paris, France) for complementary analyses that included re-isolation on Tinsdale medium, *tox* gene and species-specific *rpoB* real-time PCR,[19] Illumina genomic sequencing, biovar identification, antimicrobial susceptibility (disk diffusion, E-test; interpreted using EUCAST/CA-SFM 2019 version 2.0 if available), and the Elek diphtheria toxin production test (appendix 2 pp 2–3). Phylogenetic analyses, substitution rate, and node date estimations were based on parsnp version 1.2, SAM2MSA version 0.2.2.1, and IQTREE version 2.0.4 tools, and pangenome construction on PPanGGOLiN version 1.1.72; and known virulence and resistance genes were searched in the assembled genomes (appendix 2 pp 4–5). Sublineages were defined as deep tree branches. For the combined phylogenetic analysis of the Yemen and DIFT045 isolates with global *C diphtheriae* genomes, we used all genomic sequence assemblies that were publicly available from the National Center for Biotechnology Information database on June 1, 2020.

Statistical analysis

We tested differences in the temporal and geographical distributions of probable and confirmed cases using Fisher's exact test due to low sampling numbers in some governorates. Using a line list of patients from the first two periods, we computed the proportion of cases by age group. We calculated 95% CIs for these proportions using a binomial test. We tested differences among the proportions of cases within each age group between the two first periods using a χ^2 test. The line list was not available for the third period. Differences in clinical characteristics between cases infected with *C diphtheriae* sublineage A and those infected with other sublineages were assessed by the Mann-Whitney *U* test, the Student's *t* test, the χ^2 test, and Fisher's exact test when appropriate (eg, Fisher's exact test was used if the sample size within a category was <5). Statistical analysis was done using SPSS version 25. The diphtheria attack rate was calculated by dividing the number of cases in each age group by the size of the population within each age group (age data were obtained from the Yemen nutrition smart survey, 2019).

For the **Yemen nutrition smart survey** see https://data.humdata.org/dataset/yemen-nutrition-smart-survey

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

See **Online** for appendix 2

Results

Large numbers of probable cases of diphtheria were recorded in Yemen in 2017 (n=442), 2018 (n=2606), 2019 (n=2004), and 2020 (up to week 17: n=649). Altogether, 5701 cases, including 330 deaths, were reported up to April 26, 2020, corresponding to a case fatality rate of 5·8%. Between February, 2018, and November, 2019, we collected clinical data for 888 probable cases with throat samples referred to the NCPHL for biological confirmation (table), and genomic data for 42 positive cases corresponding to 43 isolates (figure 1). These 43 isolates were derived from 98 randomly selected samples, out of the 340 biologically confirmed cases. Some clinical data were missing (table).

We divided the epidemic into three time periods (figure 2A) and qualitatively observed incidence valleys in June, 2018, and June, 2019. In the third period

(June 4, 2019, to April 26, 2020), the recorded number of probable cases peaked in November, 2019, with more than 80 cases per week at that time. The diphtheria attack rates were 6·34 probable cases per 100 000 individuals in the first period and 6·24 probable cases per 100 000 in the second period (p=0·66).

The outbreak started in October, 2017, in Ibb and affected almost all governorates of Yemen (figure 2, appendix 2 pp 13–14, 22). The two first periods had distinct geographical distributions (p=0·0005).

The percentage of the 888 patients for whom clinical data were available who had been vaccinated did not change between 2018 and 2019 (57% *vs* 59%, p=0·49). The median age of vaccinated patients was lower (10 years, range 0·2–45) than that of non-vaccinated patients (13 years, range 0·2–80; p<0·0001). Most probable cases were in children aged 5–15 years (median 12 years, range 0·2–80), and their age distribution differed between the two first epidemic periods (p<0·0001). Specifically, there was a marked reduction in the proportion of cases that were in individuals aged 0–4 years (from 19% [95% CI 18–21] in the first period to 14% [12–15] in the second period, p<0·0001; appendix 2 pp 15–16). No difference in disease expression (characteristics in table) was observed between vaccinated and non-vaccinated patients (p values ranging from 0·12 to 0·81).

Microbiological confirmation of cases was tested for 888 samples referred to the NCPHL from 2018 to 2020. Of the 836 cases with available results, 340 (41%) were microbiologically confirmed by a positive *C diphtheriae* culture. Of note, most patients (634 [80%] of 790) were reported to be receiving antibiotic therapy (mostly penicillin, erythromycin, or azithromycin) at the time of throat sampling (78% of microbiologically confirmed cases *vs* 82% of not confirmed; p=0·23). There was no difference in the geographical distributions of probable and confirmed cases (first period: p=0·84, second period: p=0·22).

Of 98 cultures transferred to Institut Pasteur, 47 were confirmed to contain genetic material of *C diphtheriae* by real-time PCR, among which 42 could be re-isolated. YEM0065 and YEM0070 with distinct colony morphologies were isolated from a single patient. Detection of the diphtheria toxin *tox* gene was positive in 39 (91%) of 43 isolates; four isolates were *tox*-negative by real-time PCR (appendix 2 pp 6–8), which was confirmed by genomic analyses. All *tox*-positive isolates produced the toxin. Phylogenomic analyses identified six sublineages (labelled A to F; figure 3). Sublineage A comprised

	Number of cases with available data*	Characteristic, median (range) or n (%)	p value (confirmed *vs* non-confirmed cases)	Cases analysed by whole-genome sequencing		
				Cases from sublineage A (n=28)†, median (range) or n (%)	Cases from sublineages B–F (n=14)†, median (range) or n (%)	p value (sublineage A *vs* sublineages B–F)
Age, years						
Probable cases	824	12 (0·2–80)	..	..	..	..
Not confirmed	463	11 (0·2–80)	..	..	..	..
Confirmed	313	12 (0·3–60)	0·36	8 (2–35)	14 (2–35)	0·060
Cases per age group: 0–4 years; 5–14 years; ≥15 years						
Probable cases	824	151 (18%); 346 (42%); 327 (40%)	..	..	..	..
Not confirmed	463	96 (21%); 187 (40%); 180 (39%)	..	..	..	..
Confirmed	313	48 (15%); 141 (45%); 124 (40%)	0·14	6 (21%); 17 (61%); 5 (18%)	1 (8%); 5 (42%); 6 (50%)‡	0·15
Male sex						
Probable cases	881	398 (45%)	..	..	..	..
Not confirmed	496	242 (49%)	..	..	..	..
Confirmed	340	142 (42%)	0·053	14 (50%)	5 (36%)	0·51
Female sex						
Probable cases	881	483 (55%)	..	..	..	..
Not confirmed	496	254 (51%)	..	..	..	..
Confirmed	340	198 (58%)	0·053	14 (50%)	9 (64%)	0·51
History of previous vaccination						
Probable cases	750	434 (58%)	..	..	..	..
Not confirmed	427	259 (61%)	..	..	..	..
Confirmed	281	153 (54%)	0·12	11/25 (44%)	7/11 (64%)	0·47

(Table continues on next page)

	Number of cases with available data*	Characteristic, median (range) or n (%)	p value (confirmed *vs* non-confirmed cases)	Cases analysed by whole-genome sequencing		
				Cases from sublineage A (n=28)†, median (range) or n (%)	Cases from sublineages B–F (n=14)†, median (range) or n (%)	p value (sublineage A *vs* sublineages B–F)
(Continued from previous page)						
History of antibiotic therapy						
Probable cases	790	634 (80%)	..	..	..	..
Not confirmed	455	371 (82%)	..	..	..	..
Confirmed	296	230 (78%)	0·23	13/27 (48%)	11/13 (85%)	0·040
Fever						
Probable cases	841	818 (97%)	..	..	..	..
Not confirmed	476	463 (97%)	..	..	..	..
Confirmed	322	314 (98%)	0·83	28 (100%)	14 (100%)	NA
Pseudomembrane						
Probable cases	839	787 (94%)	..	..	..	..
Not confirmed	477	453 (95%)	..	..	..	..
Confirmed	319	299 (94%)	0·55	28 (100%)	14 (100%)	NA
Neck swelling						
Probable cases	769	456 (59%)	..	..	..	..
Not confirmed	437	263 (60%)	..	..	..	..
Confirmed	297	181 (61%)	0·89	22/26 (85%)	4/12 (33%)	0·0029
Laryngitis						
Probable cases	784	548 (70%)	..	..	..	..
Not confirmed	444	311 (70%)	..	..	..	..
Confirmed	299	213 (71%)	0·79	22/26 (85%)	7/12 (58%)	0·11
Tonsillitis						
Probable cases	836	821 (98%)	..	..	..	..
Not confirmed	475	468 (99%)	..	..	..	..
Confirmed	319	313 (98%)	0·87	28 (100%)	14 (100%)	NA
Difficulty swallowing						
Probable cases	823	708 (86%)	..	..	..	..
Not confirmed	468	407 (87%)	..	..	..	..
Confirmed	315	267 (85%)	0·44	25/26 (96%)	11/13 (85%)	0·25
Difficulty breathing						
Probable cases	785	399 (51%)	..	..	..	..
Not confirmed	453	234 (52%)	..	..	..	..
Confirmed	295	148 (50%)	0·72	12/26 (46%)	8/13 (62%)	0·36

Undetermined cases (n=52) were excluded. NA=not applicable. *Total number of cases was 888; the difference between the number of probable cases and the total number of cases represents missing data (eg, for age: 888–824=64 cases with missing data). †For some characteristics, the number of cases is shown as n/N; the difference between N and the denominator values (28 for sublineage A, 14 for sublineages B–F) represents missing data. ‡Two data points were missing.

Table: **Clinical characteristics of probable cases of diphtheria**

30 (70%) isolates (appendix 2 pp 6–8). The six sublineages were unrelated to each other (appendix 2 p 17). Sublineages A and B were biovar gravis, whereas the others were biovar mitis (appendix 2 pp 6–8).

Sublineage A isolates were closely related to strain DIFT045 (figure 3), which was isolated in Belgium and is epidemiologically linked to Saudi Arabia.[20] Sublineage B isolates were closely related to isolates from migrants of Eritrean origins, sublineage C isolates were closely related to isolates from migrants of Somalian origins,[3] and sublineage E isolates were related to isolates from Somalian siblings (appendix 2 pp 17–18).[21] Sublineages D and F were not phylogenetically closely related to *C diphtheriae* isolates with publicly available genomic sequences. Genetic network analysis confirmed the close relatedness of the Eritrea and Belgium isolates to Yemen isolates, and showed cross-governorate transmission of sublineages within Yemen (appendix 2 p 18).

We defined the pangenome diversity of isolates (appendix 2 p 19), identifying 3619 gene families,

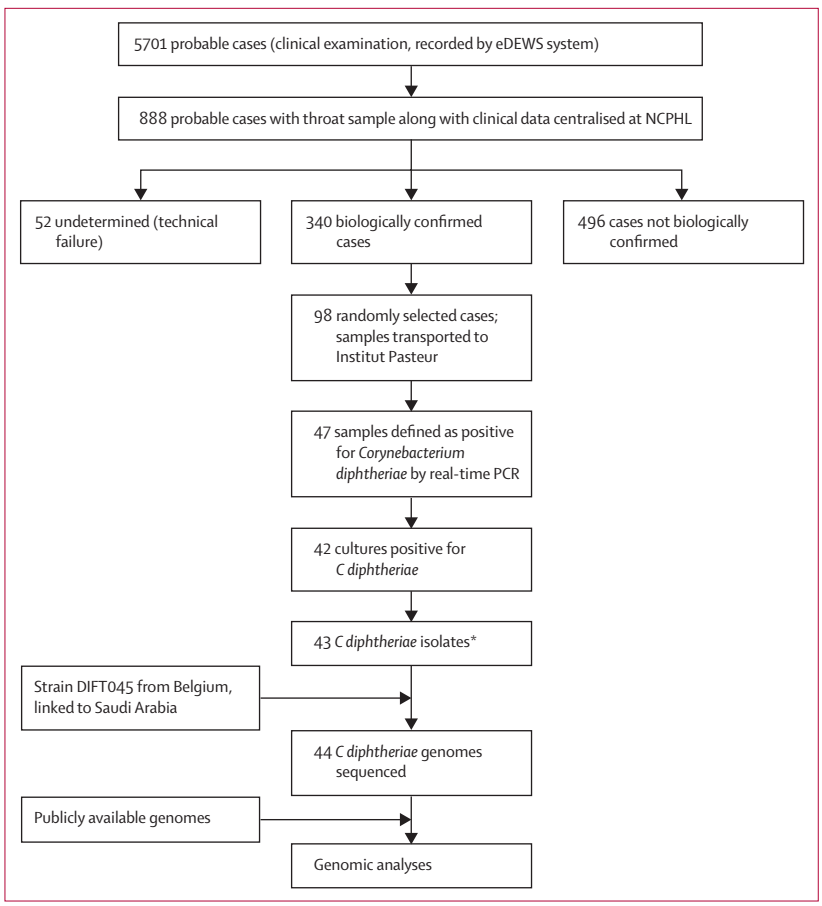

Figure 1: **Study flow chart**
eDEWS=electronic diseases early warning system. NCPHL=Yemen National Centre of the Public Health Laboratories. *Two isolates from one culture were included due to distinct colony morphologies.

1790 of which were shared by all isolates. Genomes had a mean of 2277 genes (range 2234–2427), and each of the six sublineages possessed 54–140 genes not observed in the other sublineages, underlying the potential for large phenotypic variation among them.

To evaluate the microevolution of sublineage A, we subdivided it into two branches, with branch A.1 being further subdivided into A.1.1 and A.1.2 (figure 3). The epidemiologically predominant branch A.1 was highly homogeneous genetically, with a maximum of 86 genome-wide single nucleotide polymorphisms among its members (observed between YEM0063 and YEM0089), indicating that A.1 isolates share a recent common ancestor. The amount of nucleotide substitutions observed in isolates of this sublineage was significantly associated with their isolation time (appendix 2 p 20), demonstrating measurable evolutionary divergence (figure 3), with a substitution rate estimated at $1{\cdot}67\times10^{-6}$ substitutions per site per year within A.1.1 (95% CI $3{\cdot}98\times10^{-7}$ to $2{\cdot}99\times10^{-6}$). The last common ancestor of A.1.1 was estimated to have existed in March, 2015 (range November, 2003, to September, 2016).

Antimicrobial susceptibility data were heterogeneous among Yemen sublineages (figure 3, appendix 2 pp 9–12). Full susceptibility or very low levels of penicillin resistance were observed for all Yemen diphtheria isolates, which did not carry the the *pbp2m* gene encoding for penicillin-binding protein identified in *C diphtheriae* in 2020.[22] Amoxicillin and erythromycin were active against all isolates, as were many other tested molecules. Resistance to ciprofloxacin (but not to moxifloxacin) was observed in two isolates, which both had an Asp93Gly substitution in the quinolone resistance-determining region of their *gyrA* gene. Non-susceptibility to trimethoprim was found in 32 (74%) isolates, including most members of sublineage A, but none were resistant to the combination of this agent with sulfamethoxazole. Four (9%) isolates possessed the *sul1* gene, and decreased phenotypic susceptibility to sulfonamide was observed in two of these. Resistance to tetracycline was observed in 11 (26%) isolates, which all belonged to minor sublineages, whereas sublineage A was susceptible; five of these resistant isolates had known tetracycline resistance genes encoding the Tet33 efflux pump or the TetO ribosomal protection protein. Finally, the two isolates of sublineage B had an *aadA2* gene encoding for an ANT(3′) amino-glycoside modifying enzyme. This enzyme confers resistance to spectinomycin and streptomycin, which were not tested, but not to gentamicin and kanamycin, which remained active against sublineage B isolates.

Clinical features of *C diphtheriae* infections caused by isolates from sublineage A were distinctive, with patients tending to be younger (p=0·060) and more frequently presenting with a swollen neck (p=0·0029) compared with other sublineages (table). We screened the genomic sequences for the presence of putative pathogenicity-associated genes.[23] Sublineage A isolates possessed genes *iusABC* encoding for an ABC-type iron uptake system, the *chtAB* genes that are homologous to *htaAB* genes for haemin binding, a pathogenicity island of *C diphtheriae* called PICD-11,[23] comprising a putative collagen-binding protein, and the *spaABC-strA* and *spaDEF-strBC* fimbriae gene clusters (figure 3, appendix 2 p 17). In contrast to sublineage A, most other Yemen isolates did not have the aforementioned gene clusters; their genomes comprised genes *sidBA-ddpABC* for putative siderophore biosynthesis, *irp1ABCD* for a putative ABC-type iron transporter, and the *irp2ABCDEFGHI* and *irp2JKLMN* operons. Genes encoding for SpaH fimbriae were absent from all Yemen genomes. Broader putative virulence gene variation was observed at the global scale, with a marked dichotomy between the two major lineages defined by gene *spuA* (appendix 2 p 17).

Discussion

We provide an update on the epidemiological situation of the diphtheria outbreak in Yemen, as well as a clinical description of recorded probable cases and detailed information on microbiological features of *C diphtheriae*

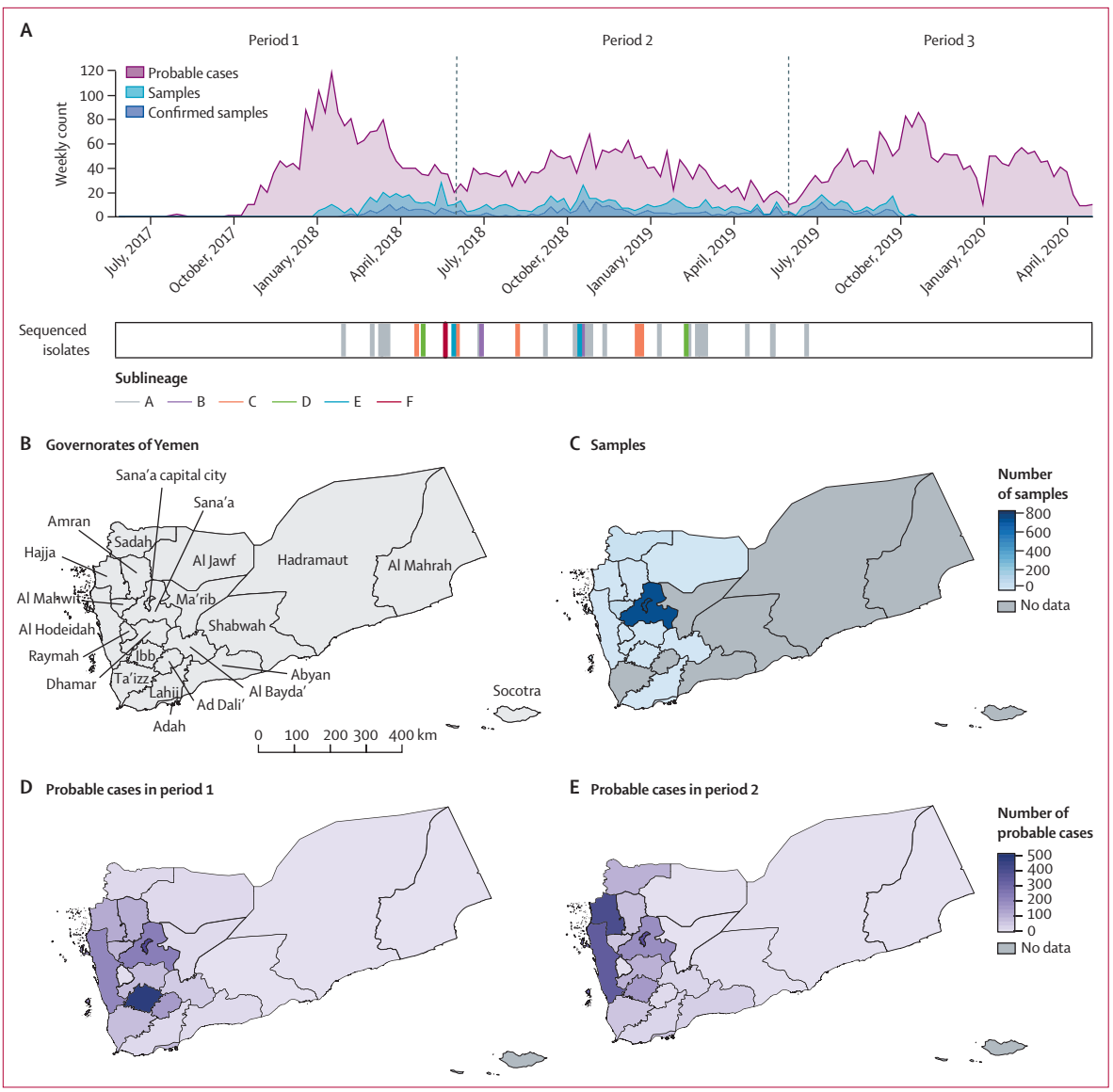

Figure 2: **Epidemic curve and geographical and temporal origins of the isolates**
(A) Weekly numbers of probable cases, probable cases with samples, and confirmed cases of diphtheria in Yemen from January, 2017, to April, 2020. Sequenced isolates are shown below the epidemic curve, coloured according to their sublineages. (B) Map of Yemen governorates. (C) Geographical location of the 888 samples of this study. (D, E) Geographical location of the suspected cases during the first and second periods. The governorate maps of Yemen were reproduced based on a shape file approved by the Humanitarian Country Team in October, 2019.

outbreak isolates. This outbreak is one of the largest contemporary *C diphtheriae* outbreaks. We defined three periods of the epidemic, which had distinct geographical distributions. Whether these periods—among which we qualitatively observed possible peaks and valleys in the number of reported probable cases—correspond to actual epidemic waves or result from variations in health-care-seeking behaviour is unclear (Ramadan and Eid occurred during May to June of 2018 and 2019). Diphtheria incidence can be affected by multiple factors, including the dynamics of internally displaced persons and active civil war.[8,9] We note that reporting is very likely to be an underestimate in the context of the Yemen health-care disruption, and to be unevenly distributed geographically, given the heterogeneity in the continuation of health-care facility operations. Nevertheless, the eDEWS system[18] was valuable in the surveillance of diphtheria in the conflict context.

Clinical characteristics were typical of diphtheria, in terms of laryngeal manifestations, pseudomembrane presence, and neck swelling. Yemenite patients were mostly young, which is also typical, with the exception of the epidemic in ex-Soviet Union countries. Vaccination is highly effective against the clinical expression of

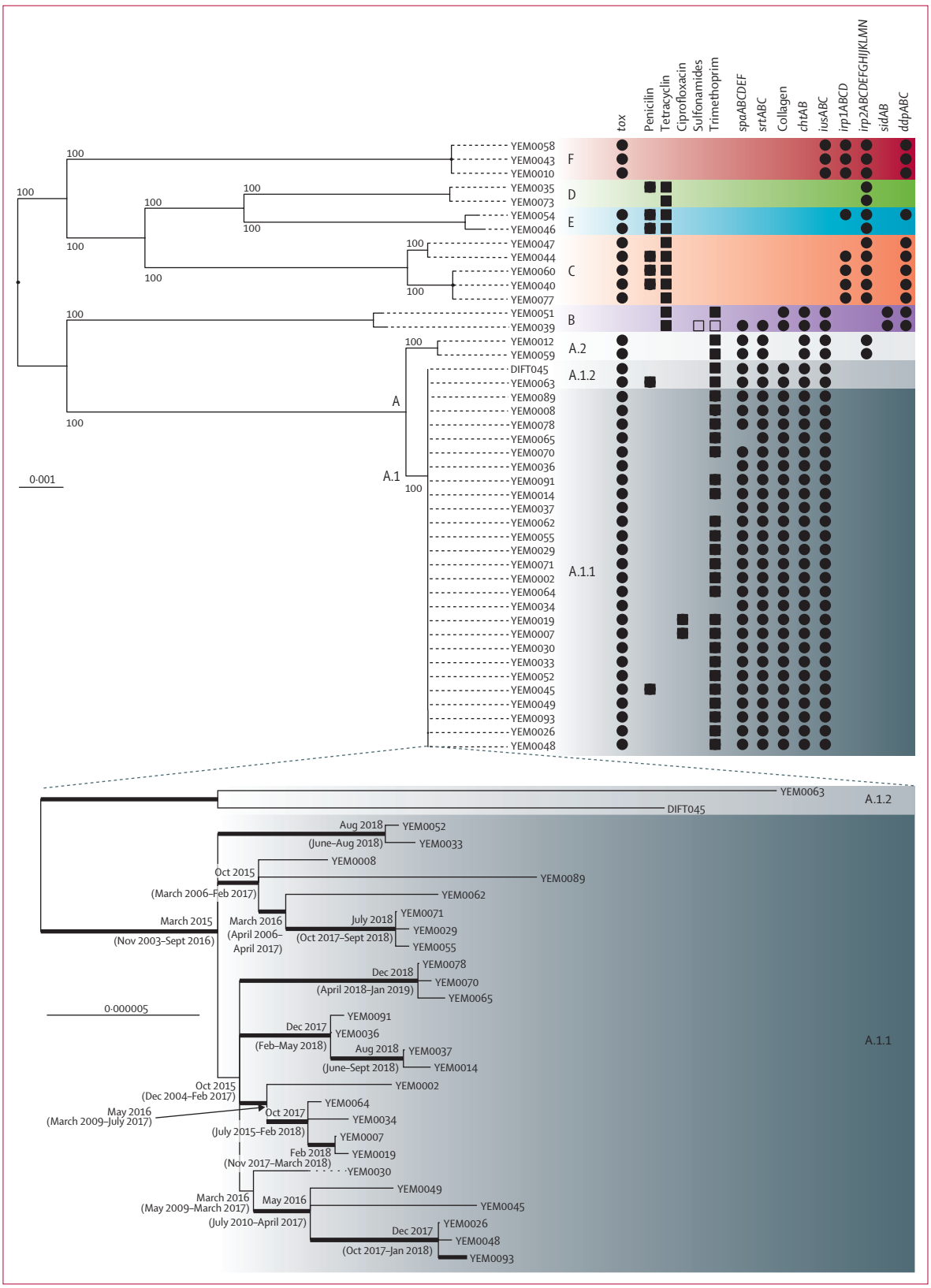

Figure 3: **Phylogenetic analysis of Yemen isolates with their virulence and resistance characteristics**
The phylogenetic tree shows 43 Yemen isolates and strain DIFT045, with the distinction of six sublineages (A to F; sublineage A was subdivided into A.1 [A.1.1, A.1.2] and A.2). Black squares denote a resistant phenotype (based on E-test for penicillin and on disk diffusion for other antimicrobial agents). Black circles denote the presence of the *tox* gene or other virulence-associated genes. Open squares denote an intermediate resistance phenotype (observed only for trimethoprim and sulfonamides). Scale bar represents 0·001 nucleotide substitutions per site. The lower part of the figure shows a detailed phylogenetic tree of sublineage A.1. Node ages (with 95% CIs) are shown at the nodes. Bold branches denote bootstrap support >80%. Scale bar represents 0·000005 nucleotide substitutions per site (ie, 1 substitution every 200 000 nucleotides, or approximately ten substitutions per genome). Isolates YEM0065 and YEM0070 were derived from a single patient sample but were both included because they had distinct colony morphologies.

toxigenic *C diphtheriae* infections, and countries with high population coverage have almost eliminated endemic diphtheria.[1] In Yemen, the vaccination coverage of the third dose of pentavalent vaccine in 2017 and 2018 was less than 80% in several governorates (appendix 2 p 21), which is considered insufficient to reach population immunity.[24] We found that more than 50% of patients were reported to be previously vaccinated. We did not have access to the exact vaccination status but can assume that some individuals had incomplete vaccination without a booster, given the high effectiveness of the complete diphtheria vaccine course. This situation might also explain the median age of patients being older than 5 years, similar to other outbreaks in populations where vaccination was incomplete.[25] From November, 2017, to March, 2018, a mass vaccination campaign targeted nearly 2·7 million children aged 6 weeks to 15 years in 11 governorates.[26] We found that from the first to the second period, there was a reduction in the proportion of cases in children aged 0–4 years, consistent with some success from the mass vaccination programme. In 2019, diphtheria vaccination was also done in 186 districts of the 12 northern governorates,[27] with 1·2 million children aged 6 weeks to 5 years receiving the pentavalent vaccine (which protects against diphtheria and four other major diseases) and 2·2 million children aged 5–15 years receiving the tetanus–diphtheria vaccine (manufactured by the Serum Institute of India). Vaccination against diphtheria in selected districts of southern governorates and Sa'adah started in July, 2020. However, despite efforts to control the disease,[27] cases were still being reported in March, 2021.[28]

This work represents a unique genomic analysis of a large outbreak of diphtheria. A prominent observation was that at least six unrelated phylogenetic sublineages contributed to the resurgence of diphtheria in Yemen, demonstrating multiple occurrences of diphtheria re-emergence or introduction. A silent diversity reservoir of *C diphtheriae* might persist despite vaccination, because vaccination is designed to protect against disease expression rather than colonisation and transmission. Antimicrobial resistance is rare in *C diphtheriae*,[29] and our results show that the main recommended agents (penicillin G, aminopenicillin, and erythromycin) are active against *C diphtheriae* isolates circulating in Yemen. However, antimicrobial susceptibility profiles were heterogeneous, with a strong phylogenetic sub-lineage effect, and a possible selection effect by previous antibiotic therapy.

Our work also uncovers a large heterogeneity in virulence-associated genomic features among circulating isolates. The diphtheria toxin-encoding gene was observed in most isolates. However, four isolates were *tox*-negative, indicating that such isolates can cause diphtheria-like respiratory symptoms. Currently, with the important exception of diphtheria toxin, the links between virulence genes of *C diphtheriae* strains and clinical expression are unknown.[30] Biovar gravis isolates have been considered more virulent than biovar mitis isolates.[31] Our genomic data suggest that inter-lineage heterogeneity in iron acquisition, adhesion, and colonisation capacities could underlie epidemiological or clinical differences. We found that patients infected with sublineage A tended to be younger and with more frequent neck swelling. However, because neck swelling is more commonly observed in young patients, we cannot separate sublineage and age effects here. The underlying mechanisms of lymph node manifestations in diphtheria, described as non-specific acute lymphadenitis,[2] are unknown. Future studies should investigate the effects of the genomic diversity of *C diphtheriae* on pathogenicity (including potential toxin expression level variation among sublineages) and clinical expression of diphtheria. The accessory genes that are unique for specific sublineages represent potential diagnostic biomarkers that could facilitate future genotype–phenotype relationship studies.

Genomic sequencing of bacterial pathogens is a powerful approach to define relationships across time, sources, and geography.[14] Migration and trade from the horn of Africa into Yemen are major drivers of pathogenic strain spread, as observed for cholera.[14] Here, two closely related *C diphtheriae* from Eritrea and Saudi Arabia were identified, and cross-governorate spread was also noted. A more systematic application of genomic sequencing would allow the geographical dynamics of *C diphtheriae* to be defined.

The estimated evolutionary rate of *C diphtheriae* is slightly faster than previous estimates[4,5] and implies that the diversification of the major Yemen outbreak sublineage largely predates the detection of the outbreak by the disease surveillance system. We note that our rate and age estimates have large confidence intervals, which is explained by the short time period and low number of available genomes.

Although we provide some insights into the microbiological characteristics of diphtheria in Yemen, the available samples of *C diphtheriae* sent for biological confirmation do not yet provide a complete picture of the outbreak. Samples referred to NCPHL were distributed across a large time period but governorates in southern Yemen were largely under-represented, because 91% of confirmed samples came from northern governorates (appendix 2 pp 13–14). This limitation reflects the current situation in Yemen, divided by conflict between the north and south. Laboratory work in Yemen yielded a low rate of biological confirmation (41%), which we attribute to operational and technical difficulties in doing microbiological analyses, including suboptimal transport and culture medium storage conditions. The high rate of antibiotic usage (penicillin or macrolides) before throat sampling could also have been a contributing factor. In addition, approximately half of the positive samples sent to Institut Pasteur were subsequently found to be negative. A likely reason is inappropriate culture conservation, because frozen storage could not be maintained due to

episodic electricity supply shortages. No check of identification was made upon subculture before shipment to Institut Pasteur, where many PCR-positive cultures were highly contaminated.

In summary, this work provides insight into the epidemiological and clinical aspects of the current Yemen diphtheria outbreak, and shows high phylogenetic, genomic, and phenotypic variation of *C diphtheriae*. Real-time characterisation of isolates at the level of individual patients could therefore be relevant for clinical management and to define strategies of vaccination catch-up or transmission tracing during diphtheria outbreaks. As such, laboratory capacity needs to be reinforced. Diphtheria is a largely forgotten disease, and present-day research into its pathophysiology is scarce. Furthermore, in the context of increasing disruption to vaccination campaigns and shortages of diphtheria antitoxin, studies into the determinants of local persistence of *C diphtheriae* and its spread at regional or global scales are needed to better control the re-emergence of this harmful pathogen.

Contributors

AA, AC-L, EB, MP-P, ML, and the NCPHL diphtheria outbreak working group were responsible for the microbiological cultures of the isolates and their biochemical and molecular characterisations. FD, GD, NL, KAAA, and HS collated and analysed the epidemiological data. AA, GD, and JT collated and analysed the clinical data. AC, VB, MH, JG, NZ, and SB analysed the genomic data. HM provided the DIFT045 strain. SB and GD designed and coordinated the project. SB designed the microbiological aspects of the study and wrote the first draft of the manuscript. AA, AC, NL, HS, JT, FD, GD, and SB had access to and verified all of the data. All authors provided input to the manuscript, reviewed the final version, and had final responsibility for the decision to submit for publication.

NCPHL diphtheria outbreak working group

Abdulaziz A Rawah, Mohammed A Aldawla, Ekram M Al-Awdi, Nabila M Al-Moalmy, Huda Z Al-Shami, Ali A Al-Somainy.

Declaration of interests

We declare no competing interests.

Data sharing

For **data from this study** see https://www.ebi.ac.uk/ena

The whole-genome sequencing data generated in this study were deposited in the European Nucleotide Archive database and are accessible through project number PRJEB34206. Raw read data are available under the accession numbers ERR4332853 to ERR4332896. Contig sequences are available under the accession numbers CAJDXH01000000 to CAJDYY01000000.

Acknowledgments

This work was supported financially by institutional funding from the NCPHL (Sana'a, Yemen) and Institut Pasteur. The French National Reference Center for Corynebacteria of the diphtheriae complex received support from Institut Pasteur and Public Health France (Santé publique France, Saint Maurice, France). This work was also supported financially by the French Government Investissement d'Avenir Programme Laboratoire d'Excellence, Integrative Biology of Emerging Infectious Diseases (ANR-10-LABX-62-IBEID). We thank François-Xavier Weill (Institut Pasteur, Paris, France) for helpful discussions, and Isabelle Cailleau (Institut Pasteur), the Médecins Sans Frontières team in Djibouti and Stephane Crisan (Epicentre, Paris, France) for logistical support. The NCPHL also acknowledges WHO for providing some reagents used in this study. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur. We also thank five anonymous reviewers who contributed significantly to improve the manuscript.

References

1 Sharma NC, Efstratiou A, Mokrousov I, Mutreja A, Das B, Ramamurthy T. Diphtheria. *Nat Rev Dis Primers* 2019; **5:** 81.
2 Hadfield TL, McEvoy P, Polotsky Y, Tzinserling VA, Yakovlev AA. The pathology of diphtheria. *J Infect Dis* 2000; **181** (suppl 1): S116–20.
3 Meinel DM, Kuehl R, Zbinden R, et al. Outbreak investigation for toxigenic *Corynebacterium diphtheriae* wound infections in refugees from northeast Africa and Syria in Switzerland and Germany by whole genome sequencing. *Clin Microbiol Infect* 2016; **22:** 1003.e1–8.
4 du Plessis M, Wolter N, Allam M, et al. Molecular characterization of *Corynebacterium diphtheriae* outbreak isolates, South Africa, March–June 2015. *Emerg Infect Dis* 2017; **23:** 1308–15.
5 Grosse-Kock S, Kolodkina V, Schwalbe EC, et al. Genomic analysis of endemic clones of toxigenic and non-toxigenic *Corynebacterium diphtheriae* in Belarus during and after the major epidemic in 1990s. *BMC Genomics* 2017; **18:** 873.
6 Wagner KS, Stickings P, White JM, et al. A review of the international issues surrounding the availability and demand for diphtheria antitoxin for therapeutic use. *Vaccine* 2009; **28:** 14–20.
7 Kneen R, Pham NG, Solomon T, et al. Penicillin vs erythromycin in the treatment of diphtheria. *Clin Infect Dis* 1998; **27:** 845–50.
8 Dureab F, Müller O, Jahn A. Resurgence of diphtheria in Yemen due to population movement. *J Travel Med* 2018; **25:** tay094.
9 Dureab F, Al-Sakkaf M, Ismail O, et al. Diphtheria outbreak in Yemen: the impact of conflict on a fragile health system. *Confl Health* 2019; **13:** 19.
10 WHO Regional Office for the Eastern Mediterranean. WHO EMRO weekly epidemiological monitor: volume 10, issue 47. Nov 19, 2017. https://reliefweb.int/report/yemen/who-emro-weekly-epidemiological-monitor-volume-10-issue-47-19-november-2017 (accessed Feb 13, 2021).
11 WHO. Diphtheria—Yemen. Disease outbreak news. Dec 22, 2017. https://www.who.int/csr/don/22-december-2017-diphtheria-yemen/en (accessed Feb 13, 2021).
12 UN Office for the Coordination of Humanitarian Affairs. Yemen: diphtheria outbreak. November, 2017. https://reliefweb.int/disaster/ep-2017-000175-yem (accessed Feb 13, 2021).
13 Meyer D. Yemen—first cholera, now diphtheria. Nov 30, 2017. https://www.outbreakobservatory.org/outbreakthursday-1/11/30/2017/yemen-first-cholera-now-diphtheria (accessed Feb 13, 2021).
14 Weill F-X, Domman D, Njamkepo E, et al. Genomic insights into the 2016-2017 cholera epidemic in Yemen. *Nature* 2019; **565:** 230–33.
15 Dureab F, Al-Awlaqi S, Jahn A. COVID-19 in Yemen: preparedness measures in a fragile state. *Lancet Public Health* 2020; **5:** e311.
16 Dhabaan GN, Al-Soneidar WA, Al-Hebshi NN. Challenges to testing COVID-19 in conflict zones: Yemen as an example. *J Glob Health* 2020; **10:** 010375.
17 WHO. Electronic integrated disease early warning and response system: volume 08, issue 19, epi week 19, (04–10 May, 2020). Geneva: World Health Organization, 2020.
18 Dureab F. The usefulness of the electronic disease early warning system (eDEWS) in the humanitarian crisis of Yemen. 2019. http://www.ub.uni-heidelberg.de/archiv/27986 (accessed Feb 13, 2021).
19 Badell E, Guillot S, Tulliez M, et al. Improved quadruplex real-time PCR assay for the diagnosis of diphtheria. *J Med Microbiol* 2019; **68:** 1455–65.
20 Martini H, Soetens O, Litt D, et al. Diphtheria in Belgium: 2010–2017. *J Med Microbiol* 2019; **68:** 1517–25.
21 Berger A, Dangel A, Schober T, et al. Whole genome sequencing suggests transmission of *Corynebacterium diphtheriae*-caused cutaneous diphtheria in two siblings, Germany, 2018. *Euro Surveill* 2019; **24:** 1800683.
22 Hennart M, Panunzi LG, Rodrigues C, et al. Population genomics and antimicrobial resistance in *Corynebacterium diphtheriae*. *Genome Med* 2020; **12:** 107.
23 Trost E, Tauch A. Comparative genomics and pathogenicity islands of *Corynebacterium diphtheriae*, *Corynebacterium ulcerans*, and *Corynebacterium pseudotuberculosis*. In: Burkovski A, ed. *Corynebacterium diphtheriae* and related toxigenic species: genomics, pathogenicity and applications. Dordrecht: Springer, 2014: 39–65.

24 Blumberg LH, Prieto MA, Diaz JV, et al. The preventable tragedy of diphtheria in the 21st century. *Int J Infect Dis* 2018; **71:** 122–23.

25 Galazka A. The changing epidemiology of diphtheria in the vaccine era. *J Infect Dis* 2000; **181** (suppl 1)**:** S2–9.

26 UN Office for the Coordination of Humanitarian Affairs. Diphtheria vaccination campaign for 2·7 million children concludes in Yemen. March 16, 2018. https://reliefweb.int/report/yemen/diphtheria-vaccination-campaign-27-million-children-concludes-yemen (accessed Feb 13, 2021).

27 WHO Regional Office for the Eastern Mediterranean. Situation report, September 2019, issue no. 9: Yemen conflict. September, 2019. https://reliefweb.int/sites/reliefweb.int/files/resources/Yem-Sitrep-Sept-2019.pdf (accessed Feb 13, 2021).

28 WHO. WHO Yemen update situation report—issue no. 3 (March 2021). May 9, 2021. https://reliefweb.int/report/yemen/who-yemen-update-situation-report-issue-no3-march-2021 (accessed May 14, 2021).

29 Zasada AA. Antimicrobial susceptibility and treatment. In: Burkovski A, ed. *Corynebacterium diphtheriae* and related toxigenic species genomics, pathogenicity and applications. Dordrecht: Springer, 2014: 239–46.

30 Sangal V, Hoskisson PA. Evolution, epidemiology and diversity of *Corynebacterium diphtheriae*: new perspectives on an old foe. *Infect Genet Evol* 2016; **43:** 364–70.

31 McLeod JW. The types mitis, intermedius and gravis of *Corynebacterium diphtheriae*: a review of observations during the past ten years. *Bacteriol Rev* 1943; **7:** 1–41.