nature communications

Article

https://doi.org/10.1038/s41467-025-60762-w

Extracting circumstances of Covid-19 transmission from free text with large language models

Received: 14 May 2024

Accepted: 3 June 2025

Published online: 01 July 2025

Check for updates

Gaston Bizel-Bizellot[1,9], Simon Galmiche[2,3,9], Benoît Lelandais[1,4,9], Tiffany Charmet[2], Laurent Coudeville[5], Arnaud Fontanet[2,6] & Christophe Zimmer[1,7,8]

Identifying the circumstances of transmission of an emerging infectious disease rapidly is central for mitigation efforts. Here, we explore how large language models (LLMs) can automatically extract such circumstances from free-text descriptions in online surveys, in the context of Covid-19. In a nationwide study conducted online in France, we enrolled 545,958 adults with recent SARS-CoV-2 infection and inquired about the circumstances of transmission in both closed-ended and open-ended questions. First, we trained a classification model based on a pretrained LLM to predict one of seven predefined infection contexts (Work, Family, Friends, Sports, Cultural, Religious, Other) from the free text in answers to open-ended questions. We achieved an unbalanced accuracy of 75%, which increased to 91% when eliminating the 43% highest entropy responses. Second, we used topic modeling to define clusters of transmission circumstances agnostically. This led to 23 clusters, which agreed with the seven predefined infection contexts, but also provided finer details on previously undefined circumstances of transmission. Our study suggests that LLM-based analysis of free text may alleviate the need for closed-ended questions in epidemiological surveys and enable insights into previously unsuspected circumstances of transmission. This approach is poised to accelerate and enrich the acquisition of epidemiological insights in future pandemics.

The Covid-19 pandemic has underscored the critical importance of swiftly identifying the circumstances of pathogen transmission to guide effective public health interventions. Epidemiological investigations aiming to determine these circumstances typically rely on interviewing infected individuals, in person, by phone, or via online questionnaires[1]. Online questionnaires, which can reach large populations at moderate cost, usually consist in series of questions targeting potential circumstances of infection. To enable statistical analyses, these questions are generally closed-ended, i.e. allow only a limited number of predefined answers. For example, questions about transportation usage may permit only the answers “car”, “bike”, “bus”, “subway” or “train”. Case-control studies containing an uninfected

[1]Institut Pasteur, Université Paris Cité, Imaging and Modeling Unit, Paris, France. [2]Institut Pasteur, Université Paris Cité, Epidemiology of Emerging Diseases Unit, Paris, France. [3]McGill University, Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada. [4]Institut Pasteur, Université Paris Cité, Image Analysis Hub, Paris, France. [5]Sanofi Vaccines, Global Medical, Lyon, France. [6]Conservatoire national des arts et métiers, PACRI Unit, Paris, France. [7]University of Würzburg, Rudolf Virchow Center for integrative and translational bioimaging, Würzburg, Germany. [8]University of Würzburg, Center for Artificial Intelligence and Data Science (CAIDAS), Würzburg, Germany. [9]These authors contributed equally: Gaston Bizel-Bizellot, Simon Galmiche, Benoît Lelandais. e-mail: arnaud.fontanet@pasteur.fr; czimmer@pasteur.fr

control population allow to identify circumstances positively or negatively associated with infection and to compute odds ratios for possible risk factors or protective factors. For instance, a recent study in France found that carpooling, attending professional gatherings, or going to bars and restaurants were associated with an increased risk of SARS-CoV-2 infection[2].

However, despite their epidemiological relevance, studies based on closed-ended questions have limitations. Questionnaires with numerous questions may discourage participants, leading to low response rates and biases in the queried population[3,4]. More fundamentally, such surveys can only characterize circumstances defined a priori as possible answers (e.g. carpooling) and are ill-suited to revealing unsuspected modes of transmission that may prevail in outbreaks of novel pathogens. An alternative to closed-ended questionnaires is to query participants using open-ended questions about their circumstances of transmission. Open-ended online queries can generate massive amounts of text information at moderate cost. However, the unstructured nature of free text is incompatible with traditional statistical analyses and hence calls for more sophisticated methods.

In recent years, the field of natural language processing (NLP) has made extraordinary progress thanks to the rapid rise of large language models (LLMs). LLMs are deep neural networks with typically ~10^8–10^{12} parameters that are trained in unsupervised ways on gigantic amounts of text data and can transform words or sentences into semantically meaningful mathematical representations (high-dimensional vectors called embeddings), in a context-aware manner. These embeddings can then be used very effectively for a wide array of tasks including translation, sentiment analysis, text generation, text summarization, or topic extraction[5–9]. Some early studies have started to explore the application of LLMs or other NLP methods to analyze free text data from web sites, questionnaires, social media feeds or electronic health records with the aim of detecting outbreaks, predicting Covid-19 case counts, detecting misinformation, improving contact tracing or public health information websites, or characterizing symptoms of infectious diseases, including Covid-19[10–20]. Thus far, however, studies leveraging NLP to identify circumstances of transmission of infectious diseases remain scarce[21–23].

Here, we showcase an LLM-based approach to identify circumstances of infection from a unique corpus of unstructured text that contains an internal ground truth, allowing us to quantitatively assess the model's predictive performance. We validate our model's capacity to quantitatively characterize predefined infection contexts and to uncover previously undefined infection circumstances from free text alone.

Results

The ComCor survey offers paired responses to closed and open questions

Between October 2020 and October 2022, as part of the ComCor study, we contacted 11,612,450 persons with a recent SARS-CoV-2 infection as determined by a positive RT-PCR test or a supervised rapid diagnostic antigenic test (i.e., performed by a healthcare worker, and not self-tests) within the past week (Methods). A total of 691,454 individuals (~6%) responded by completing an online questionnaire, which encompassed 70–100 closed-ended inquiries. Hereafter, we considered responses obtained between October 2020 and April 2022, representing a total of 545,958 cases. The questionnaire included questions on sociodemographic aspects such as age, sex, region of residence, household structure, and occupation[2,24,25] (Supplementary Table 1), and questions about workplace exposure, modes of transportation, visited locations or recreational and sporting activities. Participants were asked if they could identify the circumstances of their infection and if so to describe them, through both closed-ended questions and an open-ended question at the end. Cases who knew who infected them, or suspected a unique situation or event, were split into 3 groups: (i) cases reporting transmission through another member of their household ("intra-household", $n = 119{,}162$); (ii) cases reporting transmission from an identified person outside of their household ("extra-household", $n = 131{,}125$); (iii) cases who did not know who infected them but suspected one single transmission circumstance ("suspected situation", $n = 88{,}955$) (Fig. 1). In all three groups, participants were asked a similarly-phrased open-ended question regarding the circumstances of transmission. In the "suspected situation" group, the question (translated from French) was initially "To help us further, please give us a brief description (in a few words) of this particular event". After January 2022, the question was slightly modified as "To help us further, please give us a brief description (in a few words) of the particular situation or event", because the term "event" was found to be ambiguous in certain cases[26] (see Fig. 1 and Supplementary Fig. 1). The same group was previously asked the closed-ended question "In which context did the situation or event take place?", which allowed the following seven mutually exclusive answers: Work, Family, Friends, Sports, Cultural, Religious and Other (translated from the French "Professionnel", "Familial", "Amical", "Sportif", "Culturel", "Religieux", "Dans un autre contexte"). We focused our analyses on this specific closed-ended question and these seven infection context categories (we also use the term "transmission context" interchangeably hereafter). Please note that these seven categories were defined prior to the present study and used in previous analyses of the ComCor data set[2].

Free text about circumstances of transmission exceeds 1.3 million words

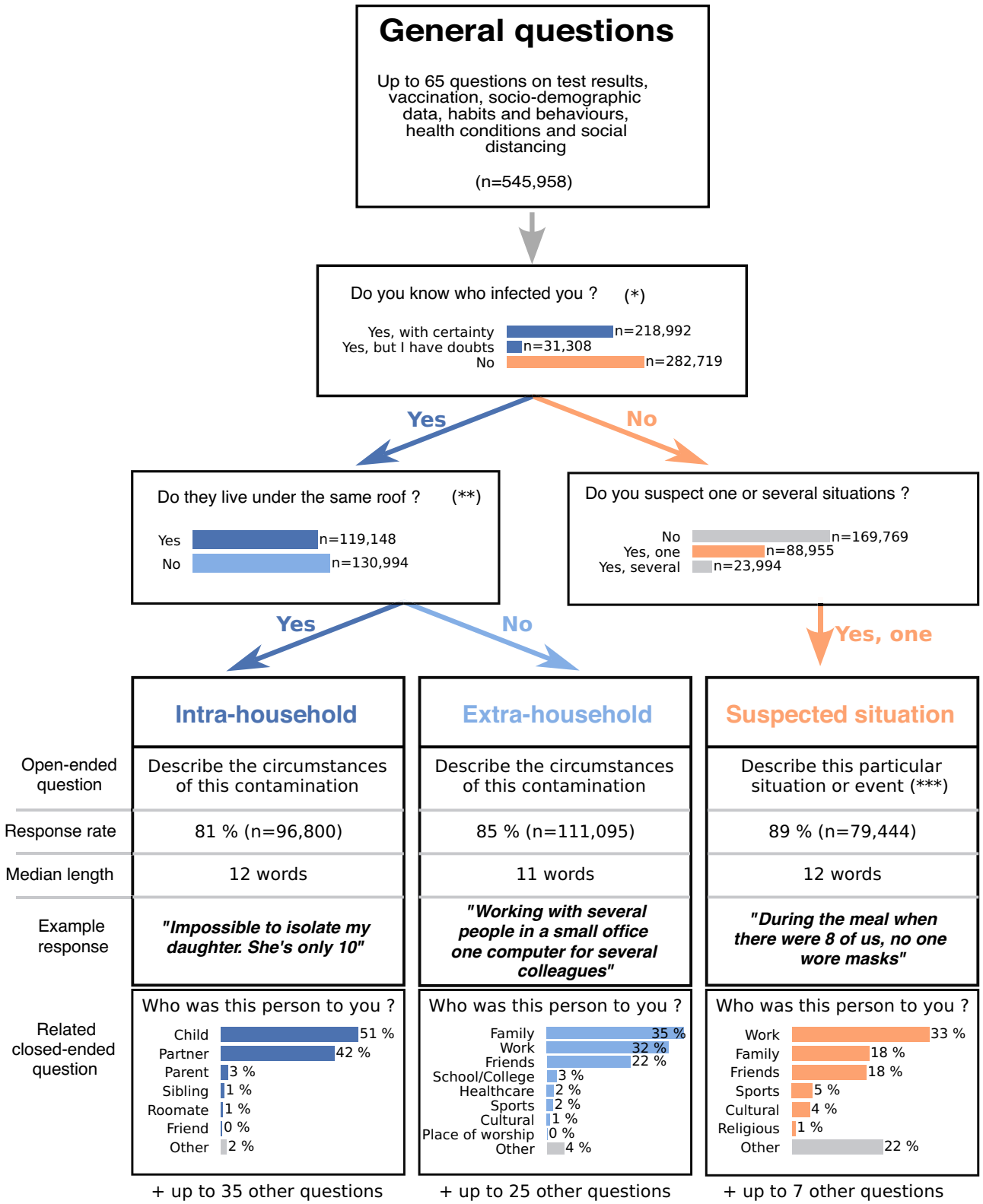
The open-ended questions led to a wide spectrum of responses, ranging from a single word (e.g. "*colleague*", "*meeting*", "*funeral*") to texts up to 1288 words long, with median and average word counts of 12 and 18.3, respectively (Supplementary Fig. 2). In aggregate, free text answers yielded a dataset of 5,247,844 words. For the group "suspected situations", the text data from $n = 79{,}444$ respondents totaled 1,349,688 words, roughly as many as in the longest novel ("A la recherche du temps perdu", by the French writer Marcel Proust, according to Guinness World Records[27]). Example answers include: "*we had a meeting in a small room where two people later tested positive*"; *"swimming pool classes"; "taking care of my grandson for a few days"; "lunch with 3 friends"; "attending a house party"; "the virus spread from someone during our ski trip because 7 out of 8 people tested positive afterwards"*. For more examples, see Supplementary Table 2. Note that all questions and answers in English mentioned in our paper were translated from French (see Methods, Supplementary Fig. 1 and Supplementary Table 2).

Training BERT to predict closed-ended question answers from free text

To test if the free text contained information that is predictive of the answers to closed-ended questions, we built a classification model that takes the raw text as input and outputs probabilities for each of the above-mentioned seven transmission contexts (Fig. 2).

For this purpose, we adopted BERT (Bidirectional Encoder Representations from Transformers)[6,8]. BERT models (and subsequent improvements including RoBERTa[8]) are LLMs that redefined the state-of-the-art for a wide-range of text processing tasks[6]. More specifically, we employed CamemBERT[28], a variation of RoBERTa designed specifically for French. CamemBERT has 110 million weights, and was pretrained on 32.7 billion tokens (a token is a unit of text usually corresponding to a word) of French text from the 138 GB size data base OSCAR[29]. OSCAR is a filtered version of Common Crawl, an archive of data crawled from billions of Internet pages. This corpus of text includes the French-speaking Wikipedia, French news sites, blogs, forums, social media platforms,

https://doi.org/10.1038/s41467-025-60762-w



e-commerce sites, educational resources and governmental sites. The CamemBERT model was supplemented with a classification head containing seven neurons, each corresponding to one of the seven infection contexts defined above. We then fine-tuned the entire model in a supervised manner, using the closed-ended answers on infection contexts as labels and a cross-entropy loss (Fig. 2 and Methods). Input text responses were truncated to the first 100 tokens before being fed to the model. We randomly partitioned the data into 72% of cases for training, 8% for validation and 20% for testing. When evaluating classification performance (see below), we considered the infection context category with the largest probability p_k, $k = 1..7$ as the single predicted context. We also calculated a normalized entropy $E = -\sum_{k=1}^{7} p_k \log(p_k) / \log(7) \in [0,1]$, which can be considered as a measure of uncertainty, since it is largest ($E = 1$) when all categories are predicted with the same probability, and is zero ($E = 0$) if a single category is predicted with probability 1.

Fig. 1 | The ComCor survey combines closed-ended question answers with free text responses to open-ended questions. In the ComCor survey, individuals tested positive for SARS-CoV-2 were asked to fill out an online questionnaire comprising a number of closed-ended questions and an open-ended question. The questionnaire begins with between 70 and 100 closed-ended questions (the number varies because some answers trigger follow-up questions that are not asked systematically). These questions concern various sociodemographic aspects (age, gender, household size, occupation, degrees,...), and aspects of potential epidemiological relevance, such as details on vaccination (number and date of shots), recent behavior (social distancing, wearing masks, smoking habits, washing hands,...) health status, symptoms, recent mode of transportation, type of accommodation, type, size and duration of gatherings (e.g. concerts) and more. Respondents were divided into 4 groups (only 3 of which are shown), depending on whether they know who infected them or not, in the first case whether or not the person who infected them lived under the same roof, and in the second case whether or not they suspect one or more situations of transmission: (i) those infected by a household member ("intra-household" cases; $n = 119{,}162$), (ii) those infected by a known person outside of the household ("extra-household" cases; $n = 131{,}125$), (iii) those who did not know who infected them but suspected one single specific situation ("suspected situation"; $n = 88{,}955$), (iv) those who did not know who infected them but suspected either no specific situation or multiple situations ($n = 193{,}777$; not shown). In each group, an open-ended question was asked, which led to free text responses as shown by the examples in bold. We focus our analyses on the $n = 79{,}444$ cases in the "suspected situation" group who provided free text responses and on the closed-ended question about the infection context on the bottom right, which allowed seven possible answers: Work, Family, Friends, Sports, Cultural, Religious, Other. Questions and responses were translated from French. See Supplementary Fig. 1 for the original French version. (*) Please note that 12,939 cases (2.4% of the total of $n = 545{,}958$) were not asked "Do you know who infected you" and hence do not show up in the downstream categories. (**) Also note that 158 individuals (<0.03%) who responded "Yes, but I have doubts" to the question "Do you know who infected you?" were not asked "Do they live under the same roof?", accounting for minor discrepancies in the reported numbers. (***) Note that the open-ended question was simplified here for brevity—see the main text for details.

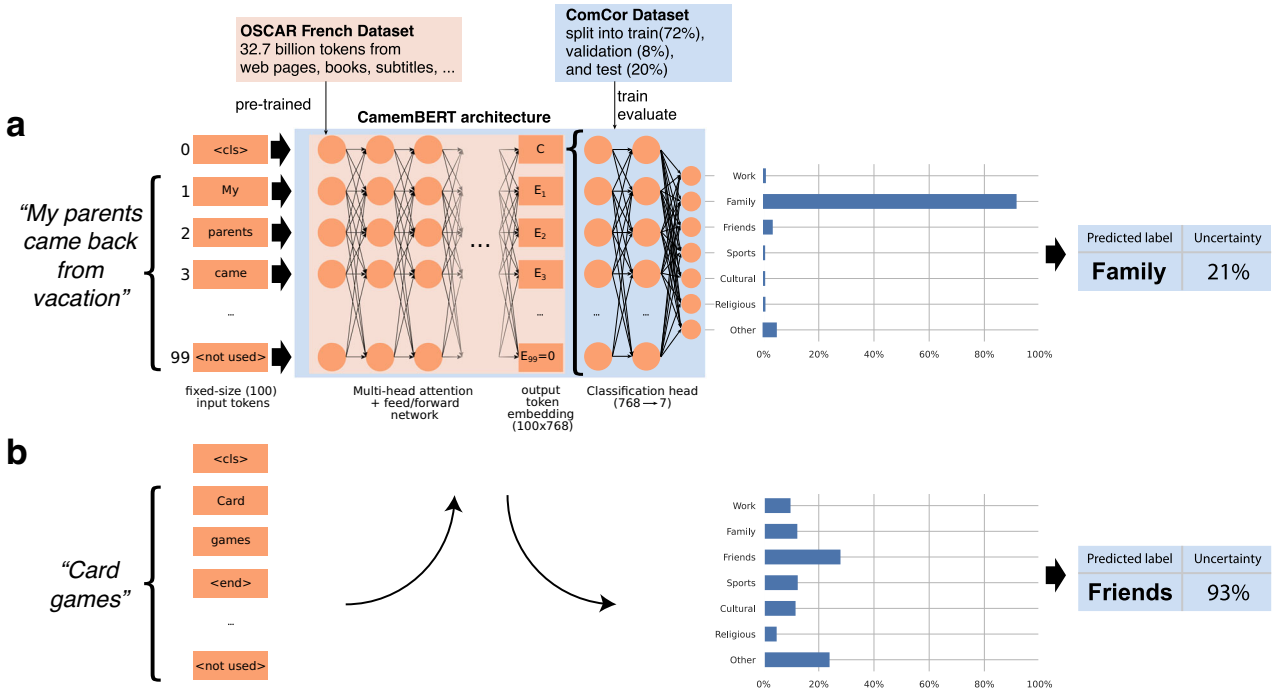


Fig. 2 | LLM based method to predict infection context from free text. Schematic of the model used to predict infection context categories from free text responses to the open-ended question. The input text (left) is split into a sequence of 100 tokens (usually: words). Longer texts are trimmed, and for shorter text, the missing tokens are padded (as indicated by the <not used> token). Each token is transformed into a 768-long vector in a context-dependent manner using CamemBERT, a transformer based LLM trained on a large French text dataset scraped from the Internet (the French version of OSCAR), consisting in 32.7 billion tokens. The embeddings obtained from CamemBERT are then fed to two fully connected layers containing 768 and 7 neurons. The last layer (with 7 neurons) has a softmax activation function to predict seven probabilities corresponding to the seven possible infection contexts (Work, Family, Friends, Sports, Cultural, Religious, Other). The entire model is finetuned in a supervised manner on free text from the ComCor survey, using as ground truth labels the context categories selected in the closed-ended question answers highlighted in Fig. 1 (bottom right). In addition, a normalized entropy is computed, which provides a measure of prediction uncertainty or ambiguity. The two examples illustrate a text with low (**a**) or high (**b**) prediction entropy.

Predefined contexts of Covid-19 infection can be predicted from free text

We aimed to predict from the free text alone the broad context of Covid-19 transmission, as defined by the seven categories Work, Family, Friends, Sports, Cultural, Religious, and Other. We recall that these seven categories were defined beforehand independently of our study[2]. In closed-ended questions, the most frequently reported transmission context was Work ($n = 5091$ out of 15,889 test cases) and the least frequent was Religious ($n = 140$). Using the test data, we quantitatively compared the context category predicted from free text to the context selected by the same individual in the closed-ended question, which we considered as ground truth. We computed balanced and unbalanced accuracies, as well as precision and recall for each context category (Methods). If the free text did not contain any information pertinent to predicting this context category, or if the model was unable to extract this information, then the classification accuracy (unbalanced) would be 21% (random classification) and the balanced accuracy 100/7 = 14.3% (Supplementary Fig. 3a). Figure 3a compares predicted categories to ground truth categories for the entire test dataset, along with precision and recall for each context

category. Overall (unbalanced) classification accuracy was 75% and the balanced accuracy 63%, which is 3.6 and 4.4 times higher than chance, respectively. This indicates that relevant epidemiological information was extracted from the free text alone, although far from perfectly when considering all cases. The model achieved highest performance for Work (precision 83%, recall 89%), while precision and recall ranged from 71% to 81% for Family, Friends and Sports and ranged from 58% to 70% for Cultural and Other. Performance was lowest for Religious, which was never predicted (recall 0%).

In order to get insights into the nature of the incorrect classifications we turned to Local Interpretable Model-Agnostic Explanations (LIME)[30] (Methods). For each word, LIME computed 7 saliency scores, each score corresponding to one of the 7 infection context categories. We then highlighted saliency scores for specific context categories in responses that were correctly or incorrectly classified (Supplementary Figs. 4a-d, 5a-d). This analysis helped explain, for example, why responses associated to the ground truth category Family were incorrectly classified as Work (Supplementary Figs. 4b, 5b). Further inspection also revealed that two of the six most frequent words for the ground truth category Religious were *"funeral"* and *"ceremony"* (Supplementary Figs. 4e, 5e). Indeed, 67% of responses associated to the choice Religious contained one of these two words. However, these two words were more abundant in the context category Family than the category Religious, since 61% of respondents whose free text contained *"funeral"* selected the context Family, and 53% of respondents whose free text contained *"ceremony"* selected Family as well (Supplementary Figs. 4f, 5f). This suggests an ambiguity or misunderstanding of these two categories among participants and points to a limitation of the questionnaire. Unsurprisingly, the model predicted the context Family for most respondents who chose Religious, with Cultural coming second (Fig. 3a).

Merging ambiguous categories based on entropy boosts accuracy

Next, we explored if classification accuracy could be improved by addressing the fact that some responses are either uninformative in themselves (such as "*I don't know*" or "*no idea*") or ambiguous for the task of predicting only one of the seven infection contexts (e.g. the sentence "*event at the opera in Switzerland with my granddaughter*" could point both to the context Cultural and the context Family). We reasoned that uninformative or ambiguous responses may be automatically identified based on the prediction uncertainty, as measured by the entropy. Indeed, the entropy of responses with incorrect predictions was on average roughly 3 times larger than the entropy of responses with correct predictions (Supplementary Fig. 6a), indicating that responses with higher entropy contribute more to classification errors. Examples of uncertain responses (entropies within the top 1%) include: "*Christmas market*", "*Book fair*", "*Sunday mass*", "*Prayers*", "*Funeral*". As illustrated by these examples, the categories Religious and Cultural were associated with the highest uncertainty (median entropies 87% and 78%, respectively), in accordance with the poorer classification performance for these categories (Fig. 3a), whereas all other categories had median entropies below 41% (average median: 30.4%) and the category Work had the lowest entropy (median 8%) (Supplementary Fig. 6b, c). On the basis of their larger prediction uncertainties, we therefore considered that the context categories Cultural and Religious were too ambiguous, and merged them with the category Other (Methods). This led to the confusion matrix shown in Supplementary Fig. 3c, with a slightly higher classification accuracy of 76.8% and a notably improved balanced accuracy of 76.0% (3.8 and 3.2 times better than chance, respectively). Thus, merging categories flagged as ambiguous by the higher prediction entropy, leads to better classification results without removing any data.

Filtering out high-entropy responses boosts accuracy

The observed correlation between entropy and prediction errors also prompted us to analyze the effect of removing the least certain predictions through increasing entropy thresholds, progressively discarding up to 95% of the test data. For the classification into seven transmission contexts (without merging categories), the (unbalanced) accuracy consistently increased and even exceeded 99% when eliminating the 90% least certain responses (Fig. 3b), still leaving $n = 1589$ test cases. However, classification accuracy tended to lose its meaning for stringent filtering, because it led to progressive elimination of all predictions for Cultural (in addition to Religious), followed by Sports, Other and Friends (Supplementary Fig. 6d, e). As a result, the balanced accuracy did not consistently improve. Instead, it remained roughly constant (in the range 58–64%) when discarding up to 43.4% of cases, but dropped to a low ~20%, when filtering out ~80% of the data. When filtering out the 43.4% least certain predictions, the balanced accuracy was 64%, almost as without any filtering, but the unbalanced accuracy was 91% (Fig. 3b, c). In this case, precision and recall ranged between 89% and 98% for Work, Family, Friends and Sports, and were 90% and 76%, respectively, for the category Other (Fig. 3c). When merging the categories Cultural and Religious with Other, as above (Supplementary Fig. 3c), the balanced accuracy was almost identical to the unbalanced accuracy and both increased consistently when filtering out up to 43% of the data. At this point, the unbalanced and balanced accuracies were 93% and 92 %, respectively (Supplementary Fig. 3d).

Thus, filtering out the least informative or most ambiguous responses using an entropy threshold allowed accurate and robust predictions of the main transmission contexts from free text alone.

Adapting topic modeling to cluster circumstances of infection

In a complementary approach, we aimed to explore whether agnostic (unsupervised) mining of the free text data can reveal epidemiologically meaningful circumstances in absence of any closed-ended questions. For this purpose, we adopted topic modeling[31], an approach that automatically partitions collections of documents into semantically related groups (topics). Specifically, we used BERTopic[9], a recent topic modeling method that -unlike bag-of-word methods[32]- accounts for the semantic relations between words in their context (Methods). To achieve this, BERTopic uses Sentence-BERT[33] (SBERT), a modification of (Camem)BERT that computes embeddings for entire sentences, or sets of sentences, rather than from individual tokens. SBERT encapsulates an entire text into a single 768-dimensional vector in such a way that semantically related texts have close embeddings. BERTopic leverages UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction)[34] to reduce embedding dimensionality, then defines nested clusters with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)[35]. HDBSCAN allows for outliers, a useful feature when handling documents dominated by "noise", which cannot be reliably assigned to any specific cluster. Finally, BERTopic uses a variation of Term Frequency Inverse Document Frequency (TF-IDF[36]) to automatically label each topic, i.e. to assign specific names (in our case consisting in two words).

Topic modeling agnostically determines circumstances of transmission from free text

We proceeded to test the potential of BERTopic to agnostically determine circumstances of infection from free text responses alone, without any training on closed-ended question data (Fig. 4). When applied to the entire dataset of $n = 79{,}444$ responses, BERTopic identified 43% of them as outliers, reflecting the relatively high noise in the data (Fig. 4a and Supplementary Fig. 7a). Moreover, BERTopic identified nine clusters (comprising 17% of responses) with names such as "*mask, wear*", "*positives, positive*", "*mask, covid*", "*pass, sanitary*", (translated from the French: "*masque, port*", "*positifs, positif*", "*masque, covid*", "*pass, sanitaire*", respectively) etc., which contained responses describing

https://doi.org/10.1038/s41467-025-60762-w

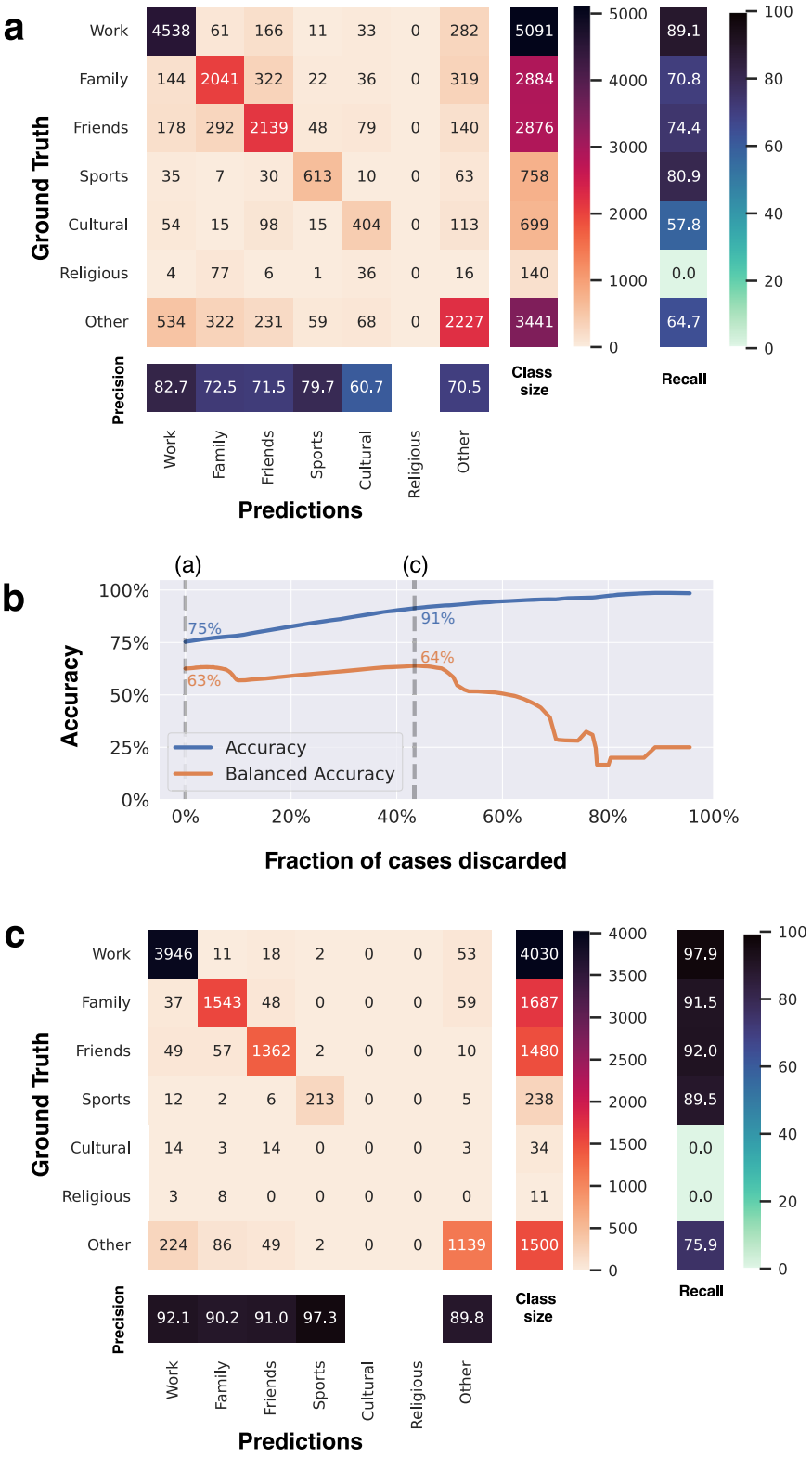
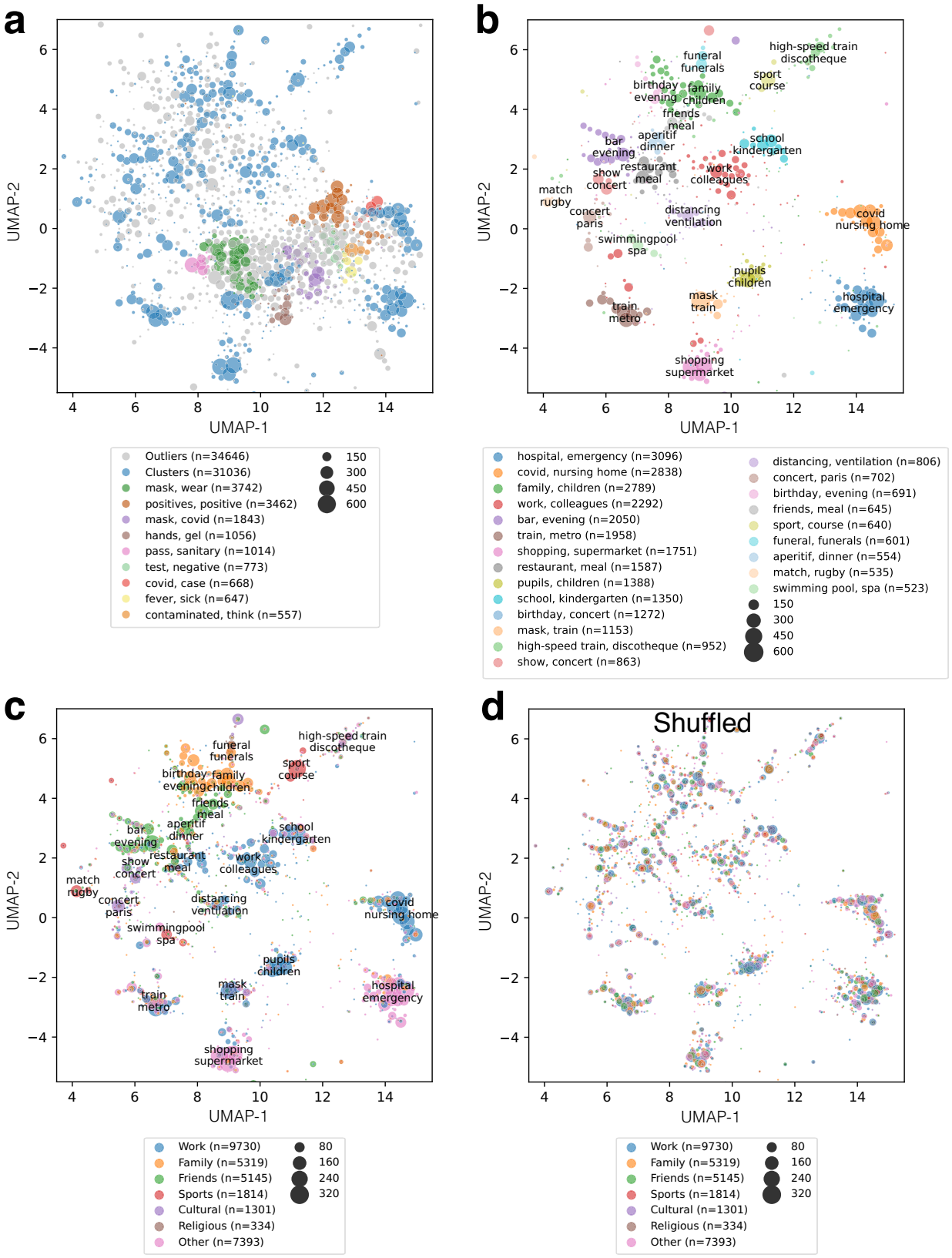


Fig. 3 | Quantitative assessment of infection context predictions. a Confusion matrix compares the infection context categories predicted by the model on the test data to the ground truth categories. Each matrix entry designates the number of cases for each pair of predicted and ground truth infection context. Numbers on the diagonal indicate correct predictions, off-diagonal entries are incorrect predictions. The total number of cases for each ground truth context is shown on the right of the matrix ("class size"). Also shown are the precisions and recalls for each context category. The unbalanced accuracy is 75.3% and the balanced accuracy is 62.5%. **b** Orange and blue curves show the balanced and unbalanced prediction accuracies, respectively, as function of the percentage of least certain (i.e., highest entropy) predictions discarded from the test data. Dashed gray lines indicate local maxima of the balanced accuracy. **c** Same as (**a**) but after discarding 43.4% of cases with the highest entropy (corresponding to the second maximum of the balanced accuracy, at 63.9%). The unbalanced accuracy is 91.3%. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-025-60762-w



situations in which social distancing was not respected. Because these responses could not be easily assigned to a more specific circumstance of infection, we hereafter ignored them and focused on the remaining 39% of responses ($n = 31{,}036$). These fell into 23 automatically determined clusters, with sizes ranging from $n = 523$ to $n = 3096$ cases (Fig. 4b and Supplementary Fig. 7b). The largest clusters were labeled *"hospital, emergency"* and *"covid, nursing home"* (translated from *"hôpital, urgences"* and *"covid, ehpad"*, respectively), and occupied a very distinct region of latent space. They consisted primarily of individuals who suspected having been infected in hospitals or nursing homes, respectively. Among the 21 remaining clusters, many were also well-defined and clearly separated from other clusters. These include the clusters

Fig. 4 | Unsupervised clustering of free text responses is consistent with closed-ended answers and provides fine-grained description of infection circumstances. This Figure shows the application of an unsupervised method (topic modeling) to determine the main circumstances of infection from free text responses only, without relying on closed-ended question answers, and visualizes its consistency with the predefined infection contexts from closed-ended question answers. **a**–**d** The plots show UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) representations of the embeddings computed by CamemBERT. Each of the small dots corresponds to a distinct text response, i.e. a distinct individual. Disks of larger size represent groups of dots belonging to the same label (color) in close proximity (distance < 0.2), with disk size indicating the number of points as per the legend. Proximity between dots or disks means that the corresponding embeddings are close to each other, indicating semantic similarity of the corresponding text responses. **a** UMAP for the entire dataset of $n = 79{,}444$ responses (but restricted to a region occupying 93.3% of the responses; see Supplementary Fig. 7a for a larger view). Three groups automatically defined by BERTopic are shown: (i) outliers (gray dots; 44% of responses); (ii) a set of 9 clusters with names such as "*mask, covid*", "*hands, gel*", "*test, negative*" (colored dots except gray and blue dots), representing 17% of responses, which did not appear to contain specific circumstances of infection but rather aggregated all responses reporting generic aspects such as a lack of social distancing, test results or health state; and (iii) all other responses (blue dots, 39% of responses). **b** UMAP showing only the latter group of $n = 31{,}036$ responses, which BERTopic partitioned into 23 distinct clusters (shown in blue under the name 'Clusters' in panel a). Here, each color corresponds to a distinct cluster, as indicated in the legend, with the number of responses in each cluster as indicated. Each cluster is automatically named (labelled) using the two most salient words for each cluster, as determined by TF-IDF. "*Nursing home*" is our manual translation of "*ehpad*", a term that was not translated by DeepL and which stands for "établissements d'hébergement pour personnes âgées dépendantes" (residential facilities for dependent elderly people). (*) The cluster "*birthday, concert*" is outside the displayed region. See Supplementary Fig. 7b. **c** Same as (**b**), except that the dots are colored according to the context of infection selected by the same individual among the seven predefined categories in the closed-ended question (Work, Family, Friends, Sports, Cultural, Religious and Other). **d** Same as (**c**), but with random shuffling of the seven context categories and without the cluster names. Source data are provided as a Source Data file.

"work, colleagues", *"train, metro"*, *"shopping, supermarket"* (4th, 6th and 7th largest clusters, respectively) as well as smaller clusters such as *"sport, course"*, *"match, rugby"* or *"swimming pool, spa"* (19th, 22th and 23th largest, respectively). Other clusters were close to each other and more consistent with a continuum of transmission circumstances. For example, the cluster *"bar, evening"*, corresponding to individuals who suspected having been infected during a festive event, was surrounded by the clusters *"aperitif, dinner"*, *"restaurant, meal"*, and *"show, concert"*. We note that BERTopic yielded less consistent labels for the two clusters "*birthday, concert*" and "*high-speed train, discotheque*", likely because the majority of the responses in these clusters contained only a single word (the median number of words within each of the two clusters was 1, but ranged from 5 to 49 for all other clusters). Such inconsistencies could presumably be eliminated by restricting the analysis to responses exceeding a minimum number of words.

Overall, despite these exceptions, the automatically defined clusters aptly captured several well-defined locations (e.g. trains, supermarkets, schools, restaurants, hospitals) and activities (e.g. work, rugby, funeral, shopping, concert). In many cases, these circumstances were not predefined in closed-ended questions (e.g. school, rugby, spa, emergency room) but nonetheless allowed straightforward epidemiological interpretation. We note that more clusters can be obtained by changing the hyperparameters of BERTopic (see Methods), potentially enabling even finer-grained descriptions of circumstances of contamination.

Agnostic clusters of transmission circumstances agree with predefined infection contexts

Subsequently, we examined the relation between these 23 clusters and the seven predefined transmission contexts discussed above (Fig. 4c, Supplementary Fig. 8). If the agnostically determined BERTopic clusters were entirely independent of these contexts, they would be distributed in the same proportions among the seven context categories and vice-versa, as shown in Fig. 4d (see also Supplementary Fig. 8d). Instead, we observed a highly non-random association between agnostic clusters and predefined contexts, characterized in most cases by the predominance of a single context for each cluster, as reflected by the relatively monochromatic appearance of most clusters in Fig. 4c (see also contingency tables and related quantifications in Supplementary Fig. 8). Indeed, a single context category corresponded to more than 75% of responses in 9 clusters, to more than 50% in 13 clusters, and to more than 40% in 20 out of the 23 clusters (Supplementary Fig. 8a–c). Interestingly, some predefined context categories were split into multiple clusters. For example, the context Work dominated the cluster *"work, colleagues"* (unsurprisingly), but also the clusters *"covid, nursing home"*, *"pupils, children"*, *"train, metro"*, and *"school, kindergarten"*. Closer inspection of the text responses suggests that the latter four clusters reflected occupational transmissions for healthcare workers, schoolteachers, commuting for work, and students, respectively, thus defining four specific professional circumstances that were absent from the predefined closed-question answers. The clusters *"shopping, supermarket"* and *"hospital, emergency"* were dominated by the context Other, thereby revealing two very distinct transmission circumstances that likewise were not captured by any predefined category.

Topic modeling breaks predefined categories into data-driven clusters

Finally, we note that the topic modeling approach can also be used in conjunction with the closed-question answers to split a specific group of respondents (based on a predefined category) into automatically labeled sub-clusters. For example, we performed topic modeling on the subset of cases who selected the context Work in response to closed questions. This analysis led to 21 clusters (not counting outliers) with clear epidemiological meanings, such as "*children, school*", "*metro, train*", "*worksite, clients*", "*meal, restaurant*" or "*mask, bus*" (Fig. 5a and Supplementary Fig. 9a). Likewise, restricting topic modeling to the cases who selected the context Cultural led to 16 clusters (Fig. 5b and Supplementary Fig. 9b). Aside from a large cluster called "*mask, pass*", most clusters corresponded to clear and distinct activities or locations, including "*choir, rehearsal*" (the second largest cluster), "*theater, shows*", and "*dance, dancing*". These examples illustrate how topic modeling can be used as a complement to the closed-question analysis to break specific infection contexts into meaningful subcategories.

Discussion

We have demonstrated an LLM-based approach that can extract circumstances of infection from free text. The pairing of free-text responses with closed-ended question answers in the ComCor data set provided a ground truth that allowed us to rigorously quantify the method's performance. With automated filtering of uncertain predictions (by entropy), our supervised classification method achieved high accuracies (e.g. 91% unbalanced and 64% balanced) in distinguishing between seven broad contexts of infection (Fig. 3). Moreover, our unsupervised clustering approach, which is not contingent on any predefined answers, agnostically determined and automatically labelled 23 distinct circumstances of transmission, that (with some exceptions) are specific, easy to interpret and are consistent with what epidemiologists now know about Covid-19 transmission (Fig. 4b). These clusters largely agreed with the seven infection contexts defined a priori, but also enriched them, by offering finer-grained partitions of these contexts and by

https://doi.org/10.1038/s41467-025-60762-w

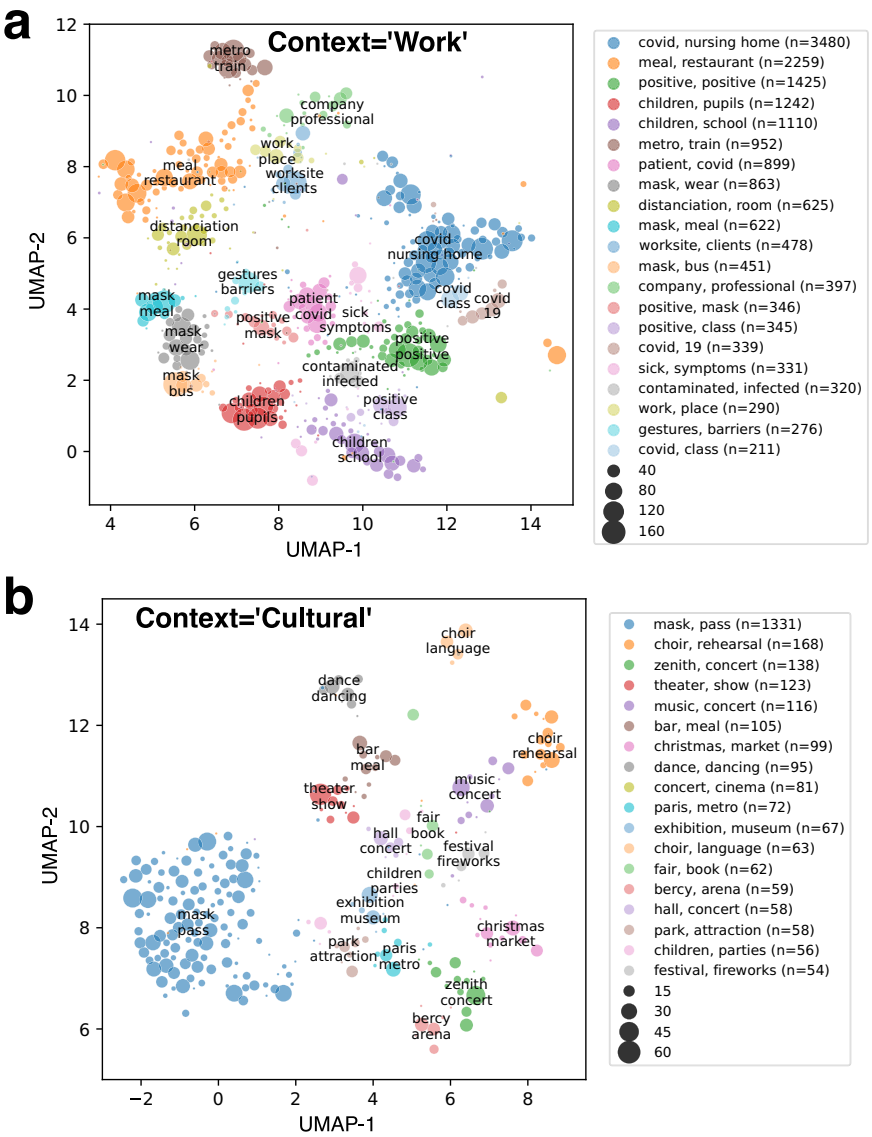


Fig. 5 | Topic modeling splits predefined infection contexts into finer circumstances. This Figure illustrates the application of topic modeling based on the free text responses to break predefined infection contexts into subcategories. Shown are UMAP representations of CamemBERT embeddings as in Fig. 4b, but restricted to the respondents who selected one specific infection context in response to the closed question, namely Work (**a**) or Cultural (**b**). Each color corresponds to a distinct BERTopic cluster as indicated in the legend. UMAPs shown in (**a**) and (**b**) are restricted to regions containing 99.7% and 97.1% of the responses, respectively. See Supplementary Fig. 9 for larger views and the original French terms. Notes on French words: "*Bercy*" and "*Zénith*" are indoor arenas and among the largest venues for sports events (Bercy), concerts (Bercy and Zénith) and more. Source data are provided as a Source Data file.

highlighting additional and specific circumstances outside the predefined categories (Fig. 4c). We also illustrated how the unsupervised analysis can be used as a complement to the closed-ended questions to automatically break down predefined epidemiological categories into meaningful data-driven subcategories (Fig. 5).

This study allows two main conclusions. First, our supervised classification and its quantitative validation show that ComCor's free text answers are useful to determine the broad context of infection. This is a non-trivial result, since respondents could conceivably omit to provide information through free text that they already provided by answering closed-ended questions ("anti-redundancy bias"). Second, it shows that pretrained LLMs are well suited to mining free text for epidemiologically relevant information and for highlighting circumstances of infection that were not defined a priori. This is a potentially important benefit for epidemics driven by new pathogens with unsuspected modes of transmission.

Several perspectives emerge. First, we propose that in future epidemics, unsupervised analysis of free-text answers with LLMs will facilitate the early identification of specific circumstances of transmission in a manner unbiased by epidemiological assumptions. Subsequently, such circumstances could be targeted by specific closed-ended questions, via rapid updates of online surveys, and thereby enable quantification of risk factors through odds ratios.

Second, the fact that answers to closed-ended questions can be predicted with high accuracy from free text opens up the possibility that these closed-ended questions could be partly or entirely dispensed of. Diminishing the number of closed-ended questions could increase the survey response rate[3,4]. An appealing possibility is to start surveys with open-ended inquiries and to follow up with closed-ended questions defined by a real-time analysis of the free text, e.g. by asking more questions when text responses have high entropy. We propose that this approach could play a role somewhat similar to that of focus group interviews[37], but with the superior speed, cost and scalability of online surveys. Asking open-ended questions earlier in the survey may also increase response quality and reduce potential anti-redundancy biases. A related option is to assign closed-ended questions to a small

random subset of individuals in order to train a model to answer specific questions from free text or to check the consistency of unsupervised topic extraction (as in Fig. 4c and Supplementary Fig. 8). This approach would combine the advantages of a light-weight free text survey with high response rate and the ground truth data provided by closed-ended questions. Entropy may also be used to flag ambiguous questions or categories (as for the contexts Cultural and Religious; see Supplementary Fig. 6b), thereby helping to redefine categories and improve questionnaires.

A third perspective, beyond surveys, is using LLMs to analyze text from news articles, social media posts, blogs, etc. Billions of individuals spontaneously share text information about their daily lives, including $> 6 \times 10^8$ daily tweets and $> 4 \times 10^9$ Facebook messages[38], without questionnaire-inherent biases. This enormous data production potentially represents an invaluable and largely untapped epidemiological resource. We suggest that LLMs may be leveraged to extract epidemiologically relevant information from these data streams in real time.

Our study has several limitations. First, responses to both open-ended and closed-ended questions may suffer from specific biases, demographic or otherwise[2], between the population who answered the survey and the population who did not (for example, women were overrepresented in the ComCor respondents; see Supplementary Table 1). Unlike for the estimation of odds ratios, we cannot make use of ComCor's demographically matched uninfected control group, since this group was evidently not asked questions about circumstances of infection. Normalizing for the frequency of activities such as dancing in the uninfected population could potentially allow quantitative risk estimates, but appears very challenging. Second, we aggregated data over a two-year period and did not consider time-varying factors that may have affected circumstances of infection (e.g. periods of lockdown) or perceived risks (e.g. the importance of wearing masks or ventilating indoor areas, for which public health messaging has evolved over time) and likely led to changes or biases in the text responses. Studying these time-varying effects is an interesting additional perspective. Third, most closed-ended questions, including the question about infection context, allowed only for a single answer. This mutual exclusion potentially restricted the classification performance and may help explain confusion between partly overlapping categories such as Cultural and Family. Fourth, we considered a single closed-ended question (the broadly defined infection context) and did not analyze if answers to other closed-ended questions can be predicted accurately. We also leave this question for future work. Finally, the unsupervised clustering yielded a large number of outliers and a few largely uninformative clusters about social distancing. Moreover, clusters consisting mostly of single words had inconsistent labels (e.g. "*birthday, concert*" and "*high-speed train, discotheque*"). Also, note that the result of clustering depends on hyperparameters (see Methods), which were set manually. More work is needed to better filter out uninformative data, avoid ambiguous clusters or automatically partition outliers into more informative topical groups. Improved clustering is presumably possible using more recent and powerful LLMs such as GPT-4.5, Claude 4, Llama4, or Mistral Large[7,39].

In conclusion, our study suggests that LLMs provide a promising avenue to extract pertinent information about the factors that drive the spread of infections from free text in online surveys alone. As this approach scales to large amounts of data and is agnostic to prior epidemiological assumptions, we believe that it will accelerate our understanding of epidemics caused by novel pathogens and hence play an important role in the public health response to future pandemics. More broadly, our study highlights the potential of LLMs to analyse complex text datasets despite limitations in the phrasing of questions towards inferring rich and actionable information relevant to public health.

Methods

Online survey

As part of the ComCor study, contact with SARS-CoV-2 positive individuals was established by e-mail using the nationwide database of the French health insurance system (Caisse nationale de l'assurance maladie, CNAM), which has the e-mail addresses of 55% of all insured people, representing 49% of the French population.

We obtained informed consent from all participants. The ComCor study received ethical approval from the Comité de Protection des Personnes Sud Ouest et Outre Mer (the Committee for the Protection of Persons South West and Overseas) on Sept 21, 2020. France's data protection authority Commission Nationale de l'Informatique et des Libertés (the National Commission on Informatics and Liberty), authorized the processing of ComCor study data on Oct 21, 2020. The ComCor study is registered with ClinicalTrials.gov (NCT04607941).

Translations from French to English

All analyses described in the paper were performed on the original French text. By default, the English translations of individual text responses mentioned in the text or shown in the Figures or Supplementary Tables were obtained using the free version of DeepL (www.deepl.com). In rare cases where DeepL failed to produce a reasonable English translation for single words, we modified the DeepL translation, either using alternative words suggested by DeepL or by manual correction (e.g. for "*ehpad*"). For the example texts shown in Supplementary Fig. 4a–d, we used the Python package deep-translator to translate the entire text from French to English, with no manual changes. Because automatic matching of each French word to each English word in the translated text was not always successful, we manually reassigned the saliency color coding (red-white-blue) of individual French words (see Supplementary Fig. 5a-d) to the semantically closest English word in the English version.

Retraining CamemBERT for classification

Text responses used as input to CamemBERT were truncated to the first 100 tokens. For each input token, CamemBERT outputs a 768-dimensional embedding. The first token (noted C in Fig. 2) represents the entire text response and is used for classification tasks. This embedding is fed to a fully connected layer of 768 neurons with a hyperbolic tangent activation, followed by a dropout layer and a last fully connected layer with a softmax activation and 7 neurons, each corresponding to one of the 7 predefined contexts of infection. This classification head thus adds 595,975 parameters to the 110,031,360 parameters of CamemBERT.

Classification metrics

Classification accuracy is the percentage of test cases with correctly predicted context categories (i.e., classes). For each category, precision is the percentage of cases for which this category was predicted that are indeed in this category; recall (or sensitivity) is the percentage of all cases really in this category that are correctly predicted. Balanced accuracy is the average recall over all categories and accounts for unequal numbers of cases between categories (class imbalance) by effectively giving larger weights to classification errors for the minority categories.

Classification into 5 classes

As described above, we retrained CamemBERT to predict 7 classes of infection contexts (Work, Family, Friends, Sports, Cultural, Religious, Other). Because of the ambiguities of the contexts Cultural and Religious, we also computed confusion matrices and associated classification metrics after merging the classes Cultural and Religious with Other, thus reducing the number of classes to five. However we did not retrain the model on these five classes but merged the predictions

from the model trained on seven classes. To this end, for each response in the test data set, we computed the sum of the softmax probabilities corresponding to Cultural, Religious, and Other and considered this sum as the probability for the consolidated Other class. As we did for the 7-class classification, we defined the predicted class among the five as the class with the highest probability.

LIME analysis

LIME (local interpretable model-agnostic explanations)[30] allows to estimate the influence of each word on the class predicted from a text response. LIME perturbs the original text by randomly removing words, generating multiple versions of the input. These input texts are then all fed independently to the model to obtain predicted class probabilities. Next, LIME performs a linear regression between a binary matrix indicating the presence or absence of words in perturbed texts and the predicted class probabilities. This regression allows to quantify the influence of each word on the prediction. For each class and each word, LIME provides a "saliency" score between −1 and 1 depending on whether the word is predictive (high values) or not (low values) of the class. We set the number of perturbations to 5000.

Topic modeling

Topic modeling was performed using BERTopic[9]. BERTopic uses the pretrained Sentence-BERT network for sentence embedding, followed by a UMAP for dimensionality reduction and clustering with HDBSCAN. Both UMAP and HDBSCAN contain hyper-parameters that control the number of clusters and the distribution of points in the multidimensional space and must be set manually (Figs. 4, 5, Supplementary Figs. 7, 9). The UMAP tool contains three hyper-parameters: the minimum distance between data points, the number of dimensions (after dimensionality reduction), and the number of neighbors. We set the minimum distance to 0 in order to allow points to pack together, and the number of dimensions to 5. We observed that fewer dimensions led to excessive loss of information while more dimensions made clustering more challenging. We therefore set the number of neighbors to 30 for the full data set analyzed in Fig. 4 ($n = 79{,}444$ cases), and to 10 for the analyses in Fig. 5 on the subset of cases who selected the context Work ($n = 25{,}846$, Fig. 5a) or the context Cultural ($n = 3{,}356$, Fig. 5b). HDBSCAN clustering uses two hyper-parameters: the minimum number of points per cluster and the number of neighbors for each point. We set these numbers to 500 and 10, respectively, for the full data set (Fig. 4), to 200 and 10, respectively, for the cases who selected the context Work (Fig. 5a) and to 40 and 3, respectively, for the cases who selected the context Cultural (Fig. 5b). By reducing the number of points per clusters, the data set can be split into a large number of clusters. Although we used 5-dimensional UMAPs for the BERTopic analysis itself, we subsequently reduced the dimensions to 2 for the visualizations shown in Figs. 4, 5, and Supplementary Figs. 7, 9. To keep the number of displayed items manageable and enhance readability, we merged proximal data points (distance below 0.2) belonging to the same BERTopic into disks with larger diameters for larger numbers of points, as indicated in the Figure legends.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The participant data of this study (with identifiers) are available from the Institut Pasteur subject to restrictions by request to A.F. Data were used under authorized agreement for this study by the French data protection authority Commission Nationale de l'Informatique et des Libertés (CNIL; the French National Commission on Informatics and Liberty). Access to these data would therefore require previous authorization by the CNIL. The study protocol and informed consent form (in French) will be made available upon request to A.F. The data will be available as soon as access is granted by the CNIL and for the duration authorized by the CNIL, which will determine the beginning and end date of availability for authorized researchers. A reporting summary for this article is available as a Supplementary Information file. Source data are provided with this paper.

Code availability

The Python codes used to generate the results in this paper are publicly available on Github at: https://github.com/imodpasteur/comcortxt or on Zenodo[40] at https://doi.org/10.5281/zenodo.15683658.

References

1. Cevik, M., Marcus, J. L., Buckee, C. & Smith, T. C. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission dynamics should inform policy. *Clin. Infect. Dis.* **73**, S170–S176 (2021).
2. Galmiche, S. et al. Exposures associated with SARS-CoV-2 infection in France: A nationwide online case-control study. *Lancet Reg. Health Eur.* **7**, 100148 (2021).
3. Galesic, M. & Bosnjak, M. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opin. Q.* **73**, 349–360 (2009).
4. Eisele, G. et al. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment* **29**, 136–151 (2022).
5. Min, B. et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput. Surv.* **56**, 30:1–30:40 (2023).
6. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805 (2018).
7. Touvron, H. et al. LLaMA: Open and Efficient Foundation Language Models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).
8. Liu, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at https://doi.org/10.48550/arXiv.1907.11692 (2019).
9. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Preprint at https://doi.org/10.48550/arXiv.2203.05794 (2022).
10. Elhadad, M. K., Li, K. F. & Gebali, F. An Ensemble Deep Learning Technique to Detect COVID-19 Misleading Information. in *Advances in Networked-Based Information Systems* (eds. Barolli, L., Li, K. F., Enokido, T. & Takizawa, M.) 163–175 (Springer International Publishing, Cham, 2021). https://doi.org/10.1007/978-3-030-57811-4_16.
11. Olmen, J. V., Nooten, J. V., Philips, H., Sollie, A. & Daelemans, W. Predicting COVID-19 symptoms from free text in medical records using artificial intelligence: feasibility study. *JMIR Med. Inform.* **10**, e37771 (2022).
12. Hripcsak, G. et al. Syndromic surveillance using ambulatory electronic health records. *J. Am. Med. Inform. Assoc. JAMIA* **16**, 354–361 (2009).
13. Kim, M., Chae, K., Lee, S., Jang, H.-J. & Kim, S. Automated classification of online sources for infectious disease occurrences using machine-learning-based natural language processing approaches. *Int. J. Environ. Res. Public. Health* **17**, 9467 (2020).
14. Mermin-Bunnell, K. et al. Use of natural language processing of patient-initiated electronic health record messages to identify patients with COVID-19 infection. *JAMA Netw. Open* **6**, e2322299 (2023).
15. Bondaronek, P., Papakonstantinou, T., Stefanidou, C. & Chadborn, T. User feedback on the NHS test & Trace Service during COVID-19: the use of machine learning to analyse free-text data from 37,914 England adults. *Public Health Pr.* **6**, 100401 (2023).

16. Towler, L. et al. Applying machine-learning to rapidly analyze large qualitative text datasets to inform the COVID-19 pandemic response: comparing human and machine-assisted topic analysis techniques. *Front. Public Health* **11**, 1268223 (2023).
17. Wagner, T. et al. Augmented curation of clinical notes from a massive EHR system reveals symptoms of impending COVID-19 diagnosis. *eLife* **9**, e58227 (2020).
18. Chen, Q. et al. Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing. *Annu. Rev. Biomed. Data Sci.* **4**, 313–339 (2021).
19. Shen, C. et al. Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland china: observational infoveillance study. *J. Med. Internet Res.* **22**, e19421 (2020).
20. Al-Garadi, M. A., Yang, Y.-C. & Sarker, A. The role of natural language processing during the COVID-19 pandemic: health applications, opportunities, and challenges. *Healthcare* **10**, 2270 (2022).
21. Schwartz, K. L. et al. Epidemiology, clinical characteristics, household transmission, and lethality of severe acute respiratory syndrome coronavirus-2 infection among healthcare workers in Ontario, Canada. *PloS One* **15**, e0244477 (2020).
22. Feller, D. J., Zucker, J., Yin, M. T., Gordon, P. & Elhadad, N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. *J. Acquir. Immune Defic. Syndr. 1999* **77**, 160–166 (2018).
23. Heider, P. M., Pipaliya, R. M. & Meystre, S. M. A Natural language processing tool offering data extraction for COVID-19 related information (DECOVRI). *Stud. Health Technol. Inform.* **290**, 1062–1063 (2022).
24. Charmet, T. et al. Impact of original, B.1.1.7, and B.1.351/P.1 SARS-CoV-2 lineages on vaccine effectiveness of two doses of COVID-19 mRNA vaccines: Results from a nationwide case-control study in France. *Lancet Reg. Health Eur.* **8**, 100171 (2021).
25. Grant, R. et al. Impact of SARS-CoV-2 Delta variant on incubation, transmission settings and vaccine effectiveness: Results from a nationwide case-control study in France. *Lancet Reg. Health Eur.* **13**, 100278 (2022).
26. Perrey, C. et al. Contributions of the qualitative Qualicor study embedded in a cohort study on the circumstances of SARS-CoV 2 infection in France. *Infect. Dis. Now.* **54**, 104943 (2024).
27. Proust, M. *À La Recherche Du Temps Perdu*. (Aegitas).
28. Martin, L. et al. CamemBERT: a Tasty French Language Model. https://doi.org/10.48550/ARXIV.1911.03894 (2019).
29. Suárez, P. J. O., Sagot, B. & Romary, L. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7).* (eds. Bański, P. et al.), (pp. 9–16). (Leibniz-Institut für Deutsche Sprache, 2019). https://doi.org/10.14618/ids-pub-9021.
30. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (Association for Computing Machinery, New York, NY, USA, 2016). https://doi.org/10.1145/2939672.2939778.
31. Lafferty, D. M. B., John D. Topic Models. in *Text Mining* (Chapman and Hall/CRC, 2009).
32. Wallach, H. M. Topic modeling: beyond bag-of-words. in *Proceedings of the 23rd international conference on Machine learning* 977–984 (Association for Computing Machinery, New York, NY, USA, 2006). https://doi.org/10.1145/1143844.1143967.
33. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Preprint at http://arxiv.org/abs/1908.10084 (2019).
34. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
35. Campello, R. J. G. B., Moulavi, D., Zimek, A. & Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data* **10**, 5:1–5:51 (2015).
36. Ramos, J. "Using tf-idf to determine word relevance in document queries." *Proceedings of the first instructional conference on machine learning.* Vol. 242. No. 1. (2003).
37. Rabiee, F. Focus-group interview and data analysis. *Proc. Nutr. Soc.* **63**, 655–660 (2004).
38. How Much Data is Created on the Internet Each Day? | Micro Focus Blog. https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/.
39. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Preprint at https://doi.org/10.48550/arXiv.2303.13375 (2023).
40. Lelandais, B. & Bizel-Bizellot, G. imodpasteur/comcortxt: release_0_1 (Version v0_1). *Zenodo*. https://doi.org/10.5281/zenodo.15683658 (2025)

Acknowledgements

The ComCor project was funded by Institut Pasteur, REACTing (Research, Action Emerging Infectious Diseases), and the French agency Agence nationale de recherches sur le sida et les hépatites virales—Maladies Infectieuses Emergentes. A. Fontanet's laboratory receives support from the LabEx Integrative Biology of Emerging Infectious Diseases (IBEID) (grant No. ANR-10-LABX-62-IBEID) and the INCEPTION project (grant No. PIA/ANR-16-CONV-0005) for studies on emerging viruses. G. Bizel-Bizellot was funded by a grant from SANOFI to C. Zimmer's lab. S. Galmiche was funded by the INCEPTION program (Investissement d'Avenir grant No. ANR-16-CONV-0005). T. Charmet was funded by the Fondation de France (Tous unis contre le virus alliance). C. Zimmer's team receives support from Institut Pasteur and the INCEPTION program (ANR-16-CONV-0005), from the Rudolf Virchow Center, University of Würzburg, and the Bavarian State Ministry for Science and Arts through the Distinguished Professorship Program. We thank the three anonymous reviewers for insightful comments and gratefully acknowledge the contribution of the participants of the ComCor study.

Author contributions

G.B.B. and B.L. performed analyses. S.G. contributed to the ComCor study and data interpretation. T.C. contributed to the ComCor study. L.C. contributed to project supervision. A.F. coordinated the ComCor study and contributed to project supervision. All authors contributed to the study design and/or revision of the manuscript. C.Z. supervised the project and wrote the paper with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-60762-w.

Correspondence and requests for materials should be addressed to Arnaud Fontanet or Christophe Zimmer.

Peer review information *Nature Communications* thanks Paulina Bondaronek and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.