Mining online behavioural data to measure zoonotic spillover risk

Bríon Ó Conchubhair[1], Maxwell J. Farrell[2], Rory Gibb[3], Luci Kirkpatrick[4], David Redding[5,6], Gregory F. Albery*[1]

1. School of Natural Sciences, Trinity College Dublin
2. School of Biodiversity, One Health & Veterinary Medicine; MRC-University of Glasgow Centre for Virus Research, University of Glasgow
3. Department of Genetics, Evolution and Environment, University College London, London, UK
4. School of Environmental and Natural Sciences, Bangor University, North Wales, UK
5. Science Department, Natural History Museum London
6. Department of Genetics, Ecology and Evolution, University College London

Correspondence to *gfalbery@gmail.com

Abstract

Efforts to predict zoonotic disease dynamics rarely directly quantify the fine-scale animal-human interactions that drive spillover. Public digital data (i.e., iEcology) could capture these behaviours at large scales, across time and space. Further, recent statistical advances allow us to semi-automate these processes, providing a framework for continuous exposure monitoring and risk prediction. Using online reviews describing macaque behaviour across southeast Asia, we demonstrate that iEcology provides a usable data source for risky animal-human interactions. We discuss challenges in operationalising these approaches, including ethical considerations and correcting spatial, taxonomic, and socioeconomic bias. In the Big Data era, semi-automated workflows like these could pave the way to dynamically updating risk maps for use in conservation and public health planning and interventions.

Keywords: Zoonotic Disease; Animal Behaviour; Big Data; iEcology; Digital Epidemiology; Human-Animal Contact; Behavioural Ecology; Risk Prediction

Fine-scale behaviour is a missing layer in zoonotic surveillance and prediction

As the rates of many infectious diseases rise, pressure is building to understand and predict the drivers of disease emergence in human populations [1]. Many emerging diseases originate in animals, from which they regularly “spill over” to infect humans, causing substantial human morbidity and mortality [2]. Some such pathogens then achieve onward transmission, becoming epidemic, and occasionally pandemic [3,4]. Meanwhile, and *vice versa*, humans increasingly spread our own pathogens into animals through the same encounters, producing a substantial conservation threat [5,6]. Wildlife-human encounters happen regularly, and often involve risky contexts including bites, scratches, and sneezes – many of which are extremely visible and potentially easy to detect and report [7–9]. However, despite the vital role that such “risky encounters” play in facilitating disease outbreaks, researchers in host-pathogen ecology and public health are only in the early stages of being able to identify them, map them, and use them to predict spillover at large scales [10,11]. Those public health monitoring systems that have incorporated behavior (e.g. app-based reporting of bites and dogs presenting as rabid) are focussed on single pathogens, potentially missing the wider risks of aggressive animal behavior. While on-the-ground “shoeleather epidemiology” and ethnographic-style behavioural studies often provide detailed and nuanced evidence for the risky encounters that spread a variety of given diseases, this knowledge is rarely passed on to behavioural sciences and the growing fields that specialise in mapping behaviours at large scales using human behavioural Big Data. Further, data are often collected in hyperlocal contexts, not geospatialised for privacy or other reasons, and rarely collected or measured in standardised ways that make quantification particularly feasible at scale.

Predictive frameworks for zoonotic disease risk have advanced substantially in the last few decades [12,13]. Models can use host features like biogeography or life history to estimate an animal’s probability of sourcing a pathogen [14,15] and predict the arising risk distributions [16]; they can use pathogen traits like transmission or replication mode to identify its ability to infect a host (most often humans) [17]; and genomic approaches have come so far as to predict receptor binding capacity purely from genetic sequences of hosts and pathogens [18]. Notably some of these models include behaviour-proximate traits: for example, bats’ roosting ecology is associated with their status as hosts of zoonotic viruses [19], and across mammal species, the

ability to inhabit urban environments and use human resources predicts observed pathogen richness [20]. However, due partly to differences in scale and variation in individual behavior, these proxies fall short of modelling and predicting the actual rates of wildlife-human interactions as a measure of zoonotic risk. This is a vital filter, which has previously been identified to be an important priority and a shortcoming of the current focus in the host-pathogen modelling approach [20]. Consequently, there is a growing recognition of the fact that animal-human contact should be more evenly and efficiently quantified [21].

Risky encounters can take a wide variety of shapes that vary across infectious agents and transmission modes. For example, rabies is spread through dog bites [7]; Herpes B virus can be contracted through fluid contact when feeding macaque monkeys [8]; Nipah virus from consuming palm sap contaminated with bat faeces [22]; hantaviruses from sweeping rodent droppings [23]; Ebola from butchering and carcass handling [24]; and bovine tuberculosis (*Mycobacterium bovis*) from close-contact coughs and sneezes [9]. For many of these pathogens, infectivity is high enough (and evenly distributed enough across humans) that rates of behavioural exposure are a very strong predictor of realised or observed infection rates. Indeed, the field of network epidemiology as a whole rests on the assumption that behaviours mediate encounters and therefore exposures, which drives much observed variation in infection risk [25]. Behaviour is therefore an integral component of transmission risk, and often the signal present in behaviour and contact rates is used as a direct proxy for this risk, informing policy and decision-making in human disease (e.g., in SARS-CoV-2 [26]). In contrast, we are generally able to estimate proxies for risky human-wildlife contacts only at the coarsest scales (e.g. via species range overlap [27] and human development indices [17]), with very little knowledge of how this translates to probabilities of interaction and disease transmission at highly local, epidemiologically relevant scales, and how this varies across the landscape and through time. Here, we present a case study and discuss the feasibility of developing a system to do so, using modern high-throughput methods based on publicly available online data.

Digital data sources can reveal wildlife-human interactions

Increasingly, online ("iEcology") data are being used to inform on a variety of patterns and processes in ecology and evolution, covering both fundamental and applied aspects of these fields [28,29]. These datasets are a valuable complement to traditional field survey data, with advantages including substantially lower financial and logistical costs, the ability to study wildlife in poorly monitored or highly human-dominated environments, and (most importantly) dramatically increased observational scale. In principle, collectively, these properties could allow researchers to quantify behavioural interactions across broader spatial, temporal, and taxonomic gradients, facilitating advanced comparative analyses of zoonotic risk across species, environments, and human contexts. For over a decade, digital data like these have been brought to bear widely on applied problems in public health [30–32] and conservation [33–38], but these have yet to be fully capitalised on to address the specific (but, we suggest, very tractable) problem of spillover risk.

Online data sources like these take a variety of forms, representing reports across a wide range of scenarios in which humans and animals are likely to be interacting, without participants necessarily realising they are reporting an event associated with a risk of disease transmission [28]. Because these reports include a signal of risk that is often generalisable across space and time and detectable using standardised protocols, mining these public data is likely to be highly scalable. Such data are usually amenable to automatic processing, with less subjectivity (i.e. where human data extraction varies according to who is doing it), increased replicability, and the ability to tweak, modify, update, and experiment with different approaches without repeating substantial manual effort [39]. There are well-established uses for data such as iNaturalist and GBIF, which function well for (for example) understanding where animals spend their time and how [40,41]. The iEcology data we describe have potential to improve on these data in some ways for the purposes of zoonotic disease prediction. First, people may be especially likely to report negative interactions when reviewing a location, for example [42]; second, these data often emerge specifically from human-dominated environments and wildlife-human interfaces, such as tourist attractions, resorts, urban environments, transport hubs, or food-associated contexts. In these contexts, direct exposure opportunities are likely to occur but formal ecological monitoring is often limited. Finally, (as we discuss below), these approaches may substantially increase the scale and diversity

of observable wildlife-human interactions, enabling comparative analyses of behavioural risk across species, locations, environments, and social contexts [28,43].

Individual studies have neatly demonstrated that digital traces can reveal specific human-wildlife behaviours relevant to pathogen transmission, including tourist proximity to gorillas [6] and risky practices during hunting of wild pigs [11]. However, these approaches remain largely isolated case studies: online behavioural data have yet to be systematically incorporated into zoonotic disease monitoring and prediction at large spatiotemporal scales. Why? One likely reason is that most monitoring systems remain primarily focused on pathogens themselves: identifying them, estimating their prevalence, characterising their reservoir hosts, or documenting realised spillover events [4,13]. By contrast, behavioural interactions represent a more indirect or approximate component of risk, capturing the exposure opportunities that facilitate transmission rather than the results of the transmission events themselves. In some ways, identifying exposure events, many of which do not result in disease transmission, may superficially represent a lot of wasted effort in surveillance. However, *successful* spillover events are generally rare, stochastic, and difficult to observe directly due to the multitude of filters at play [44,45]. This is particularly true in advance of outbreaks' emergence, which is when such surveillance is most fruitful (i.e., prevention is better than cure). Monitoring human-animal interactions via online data could, in principle, offer an additional proactive approach to identifying early warning signals of pathogen spillover and emergence risk – perhaps alongside other emerging approaches such as wastewater surveillance [46,47].

Indicatively, a recent outbreak of Andes hantavirus occurred on Dutch cruise ship *MV Hondius*, attracting international news attention. The outbreak was reportedly (but as yet unconfirmed) linked to tourists who had visited an Argentinian refuse heap on a birdwatching expedition, highlighting the importance of understanding zoonotic risk specifically within highly mobile human populations interacting with wildlife in recreational environments. Social media was later used in retrospect, in combination with other more classical tools, to reconstruct the pathways of emergence (Redding et al., in review). In contexts like these, online reviews and

tourism-associated digital trace data may be especially valuable precisely because they are generated by the same populations potentially experiencing exposure risk. Tourist-facing platforms may therefore provide unusually sensitive indicators of emerging behavioural risk for specific diseases at wildlife-human interfaces that are otherwise poorly captured by traditional ecological surveillance systems. They do, however, carry other shortcomings that we discuss later (see “Caveats” section).

More broadly, the value of digitally sourced data like this often lies in the fact that people inadvertently reveal information beyond the explicit purpose of their posts. For example, analyses of Amazon reviews describing scented candles as having "no smell" proved useful large-scale indicators of COVID-19-associated anosmia [48]. Similarly, written reviews, photographs, videos, and social media posts describing wildlife encounters and observations may unintentionally record behaviours that are epidemiologically informative despite the authors having no intention of reporting disease risk. Rather than treating these digital records simply as opinions or anecdotes, behavioural surveillance seeks to recover the latent ecological and epidemiological information embedded within them.

How, then, are we to use internet data to derive useful behavioural risk information? Put simply, we need to: 1) Identify behaviours that present a risk of disease transmission. 2) Detect a digital signature of these behaviours in openly available online data (for example, text posts reporting them or images depicting them). 3) Link them to locations in time and space and the species involved. 4) Integrate additional risk factors like tourism intensity, urbanisation, provisioning opportunities, waste accumulation, host density, or environmental conditions, identifying spatial and temporal hotspots of risky interactions. 5) Link rates of the outcomes (risky encounters) with their drivers across time, space, and species. Finally, we may 6) use these workflows to develop semi-automated pipelines for collecting, processing, and analysing the data, allowing us to flexibly predict disease risk in time and space.

A case study of macaque monkeys in online tourist reviews

To demonstrate the feasibility of using iEcology data to measure risky encounters across highly scalable online media, we conducted a case study on a well-studied and important genus for disease control: macaque monkeys (*Macaca* sp.). We

selected macaques as a model system because they are widespread, abundant, heavily studied, and commonly associated with close human contact, tourism, provisioning, and urban conflict [49]. They are also known to host a diversity of zoonotic pathogens [50], including (for example) the highly deadly herpes B virus [8]. We identified candidate macaque populations and wildlife-human interface sites using a selection of tourism sites known to support frequent human-macaque interactions, pairing them with nearby sites from MacaqueNet [51], an aggregated meta-database of social network data in macaque populations.

To quantify risky human-wildlife interactions, we used publicly available Google Maps reviews as the primary source of behavioural data. These reviews frequently contain detailed descriptions of wildlife encounters, ranging from highly positive to highly negative. Whether perceived as positive or negative experiences, many such encounters could have negative disease consequences, including aggressive interactions, scavenging, contamination, theft, hand feeding, and other physical proximity events. We scraped review text and associated metadata from 48 macaque-linked locations using an automated collection pipeline. From each of these locations (Figure), we extracted a random selection of up to 1000 reviews; if there were less than 1000 reviews we extracted as many as existed (27, 56%; median number of reviews per site: 731.5). These were often in English, but where they were posted in other languages, they came with Google's autotranslated version of the text (Figure).

To give an overview, many described aggressive food-stealing behaviour, being bitten, having to be treated (e.g. for rabies), and having to alter their behaviour in the wake of the encounter (e.g. not being able to swim or carry out other tourist activities), often warning visitors not to feed macaques in certain places. Together, these confirm the presence of regular risky interactions, with medical consequences, and with purported geographic variation.

To process the reviews systematically, we used simple natural language processing and keyword-based text mining to identify references to macaques and classify reviews into risk categories. Specifically, we identified reviews including terms related to monkeys/macaques, using the full list of terms in the supplement. We classified risky behaviours by searching for words associated with aggression, bites,

scratches, or food theft, classifying reviews as 0 if these were not mentioned, and 1 if they were (Figure).

For each location, we counted frequency of 1) macaque mentions, and 2) risky behaviour mentions. We discovered that reviews mentioning monkeys were substantially more likely to mention these risky interactions (Figure). We then aggregated behavioural signals by site; to reduce the influence of uneven reporting effort, we standardised behavioural mention frequencies by total review volume assessed (max 1000) to account for differences in tourism intensity and online activity (Figure). We mapped geographic heterogeneity in macaque-human interaction risk (Figure), identifying behavioural hotspots. To quantify this risk surface statistically, we fitted an integrated nested laplace approximation (INLA) model using the prevalence of aggression mentions as the response variable, with a binomial family and using the longitude and latitude of the sites to model spatial autocorrelation. This approach models two-dimensional variation in the response variable, testing whether sites that are closer together are more likely to have similar values (here, for aggression). We found detectable spatial heterogeneity, with spatial autocorrelation effects improving model fit by 3.45 ΔDIC (change in Deviance Information Criterion). This spatial heterogeneity could arise from variation in a number of factors: anthropogenic, environmental, or otherwise. For example, anecdotally, the macaques of the Thai city Lopburi had become reliant on food from the hands of tourists. When the flow of backpackers came to a halt over the lockdowns during Covid-19, this supply of food ran dry, which left the macaques “hungrier and more aggressive than before”. This then led to a substantial change in their behaviour, leaving them more likely to steal food in the wake of the pandemic [52].

Through this workflow, we developed a proof-of-concept framework for behavioural zoonotic surveillance using public digital data, demonstrating how we can transform unstructured online interaction records into spatially explicit indicators of wildlife-human contact risk. By deriving variables indicative of environmental or population- or species-level factors, subsequent steps to the modelling pipeline could link behavioural risk with its causal drivers: for example, does local tourism rate predict local macaque aggression? At what scales are these two processes linked?

(i.e., at the site, county, or country level?) Ultimately, given enough data, these could be further incorporated into (much more advanced) predictive models.

While this is a simplified version of our idealised digital behavioural risk prediction workflow, we demonstrate that even with basic NLP approaches and keyword detection, we are able to identify clear patterns in behaviour based disease risk. Expanding this approach to include more powerful language and vision models, future iterations will be able to more flexibly account for variation in semantics, behaviour types, study species, and other environmental factors such as seasonality. While there is much room to expand on this case study, we believe it functions as proof-of-concept that online media can be used to detect heterogeneous patterns in probable disease risk. Below, we address these possible challenges, which may need to be overcome to operationalise these methods, in more detail.

Caveats regarding equity and the representativeness of online monitoring

A central challenge for digital behavioural surveillance is representation. Public online content is not generated uniformly across people or places, and is likely to over-represent populations with reliable internet access, higher social media use, and a greater propensity to document and share unusual experiences online [28,43,53,54]. Consequently, these approaches are likely to preferentially capture high-intensity interactions experienced by tourists and other relatively privileged groups while under-representing the regular, day-to-day contact and sharing of space with wildlife experienced by many rural and marginalised communities, where zoonotic burden is likely to be greatest [4,55–57]. Rather than viewing this as a fatal limitation, we argue that it identifies a key research priority: determining which populations are represented (and underrepresented) by different digital data streams, quantifying these structural biases, and integrating complementary sources of behavioural information where important gaps exist.

Our macaque case study is valuable precisely because it represents a system in which this bias is likely informative. Tourist populations are both highly exposed and highly likely to document conspicuous interactions such as aggression, provisioning, or food theft, making them a useful proof of principle that behavioural risk signals can

be recovered from online data (as demonstrated with the Andes hantavirus outbreak; Redding et al., in review). They are also very likely to access medical treatment following high-risk exposure. However, extending these approaches to broader zoonotic surveillance will require deliberate efforts to improve representativeness. This may include mining posts written in local languages rather than relying primarily on international tourism platforms, incorporating additional digital sources such as local news reports, community forums, and regionally popular social media platforms, and comparing reporting patterns between tourists and local residents to explicitly estimate reporting biases [30,31]. Integrating these behavioural signals with independent information on human demographics, animal abundance, socioeconomic factors, and existing disease surveillance [57] will also help distinguish real ecological variation in contact risk from variation in who is represented online.

An additional question is the extent to which behavioural signals derived from tourist interactions might reflect broader patterns of animal behaviour – and more generally, how signals derived from one source map onto another. For example, locations where macaques are frequently reported to steal food from or behave aggressively towards tourists may also represent locations where animals exhibit greater boldness, habituation to humans, or propensity for close interactions more generally [58,59]. That is, it may be unlikely that macaques distinguish tourists from locals, and therefore risk to one correlates with risk to another. If so, tourist-derived behavioural observations may provide useful indicators of realised contact risk for local communities as well, despite differences in the populations generating the data [59]. However, cultural norms about interacting with animals can elicit different behavioural responses: for example, primates may use the presence of sun glasses and backpacks to target thefts to tourists specifically [60]. Evaluating the degree to which behavioural risk inferred from tourists predicts interactions experienced by local residents therefore represents an important avenue for future validation with this specific pipeline. Fundamentally, these approaches could be addressed by characterising the human behavioural component of spillover risk. Rather than simply documenting where wildlife-human interactions occur, future work could use the same data to ask why some people, communities, or cultures are more likely to feed wildlife, approach animals closely, tolerate their presence, or share

environments with them [61,62]. Combining behavioural surveillance with demographic, socioeconomic, cultural, and ecological information may therefore reveal the drivers of risky interactions from the human side of the equation, providing a foundation for targeted behavioural interventions and more equitable risk prediction. However, linking behavioural traces to sensitive personal or community characteristics introduces substantial ethical risks, including privacy infringement, stigmatisation, and the possibility that surveillance systems could reinforce existing inequalities [36,43].

Importantly, we do not suggest that these approaches could replace conventional surveillance or provide unbiased estimates of absolute spillover risk [63]. Rather, we suggest that they contain an informative behavioural signal that is currently almost entirely absent from operational risk quantification systems [4,57]. Even imperfect behavioural indicators could substantially improve our understanding of where and when risky wildlife-human interactions occur, particularly when integrated with existing data on host distributions, environmental conditions, and pathogen occurrence across various scales [12,27,45]. We therefore envisage digital behavioural surveillance as most immediately applicable to systems with conspicuous wildlife-human interactions, frequent public reporting, and strong behavioural heterogeneity: situations in which coarse variation in human or animal density likely do a poor job of approximating contact processes [20]. Extending these approaches to additional host-pathogen systems will require integrating additional data sources and validating which forms of digital behaviour best reflect realised exposure risk for the given system. For example, although vector-borne pathogens do not typically involve conspicuous wildlife encounters, online reviews and social media posts describing places as "full of mosquitoes" or "covered in ticks" may nevertheless provide scalable indicators of perceived vector exposure and nuisance, offering another route by which digital behavioural data could complement conventional surveillance [64,65].

Challenges for behavioural digital monitoring

Besides the representativeness of the target human group and wildlife taxon, a major challenge for behavioural surveillance using online data is participation and reporting bias. Individuals may be substantially more likely to post reviews following unusual,

emotionally salient, or negative experiences – for example, after being bitten by wildlife – which could skew the patterns of reporting present in these reviews [54]. However, this bias is not necessarily a weakness in the context of zoonotic risk surveillance: indeed, one reason we selected online review platforms as a data source is precisely because they are likely to be especially sensitive to rare but consequential aggressive interactions that are highly relevant to spillover risk for certain pathogens. Similar reporting biases have proved informative in other forms of digital epidemiology, as in the “no smell” candle reviews outlined above, where unusual consumer behaviour inadvertently revealed epidemiologically meaningful signals [48]. More broadly, digital trace data are likely shaped by multiple overlapping forms of behavioural and social bias, including tourism intensity, cultural variation in reporting behaviour, platform demographics, and species charisma [33,36,37,66]. Future work could potentially account for some of this heterogeneity by incorporating reviewer-level baselines, such as overall review sentiment or individual tendencies toward positive or negative reporting behaviour. Such approaches may help distinguish true ecological variation in risky interactions from individual-level variation in reporting style, although they would also introduce additional methodological and ethical considerations surrounding user-level data processing and privacy [67,68].

Related to the question of species’ charisma, another important limitation is observation bias in terms of taxonomy or behaviour. Many zoonotic transmission pathways involve cryptic or indirect interactions that may not produce obvious behavioural signals in online text or images. For example, exposure to hantaviruses through sweeping contaminated rodent droppings [69,70] or Nipah virus transmission through consumption of bat-contaminated date palm sap [71,72] do not require direct wildlife encounters that are easily observable or routinely documented online. However, even when the transmission behaviour itself is not directly visible, digital traces may still capture associated ecological or environmental risk signals. Reviews and social media posts may describe bats roosting near accommodation, wildlife presence around food preparation sites, poor waste management, or frequent rodent activity, all of which may indirectly indicate elevated exposure risk [4,45]. Therefore online behavioural surveillance may still provide useful information on the broader ecological contexts and contact opportunities associated with spillover, even when the actual transmission event remains unobserved.

Once the models are carried out, it may be a challenge to validate them: this will require linking predicted behavioural signatures with epidemiological outcomes derived from disease surveillance in humans. Even assuming that sufficient preemptive measures are not taken to reduce spillover rate, these relationships are likely to involve substantial temporal complexity, including delays associated with pathogen incubation periods, reporting lags, and stochasticity in spillover dynamics [4,45]. Consequently, behavioural hotspots may precede detectable epidemiological signals by days, weeks, or months depending on the host-pathogen system. Validation may become especially difficult in highly mobile human populations such as tourists, especially where infections only become clinically apparent after individuals have travelled long distances through national or international transport networks [73,74]. In such systems, spatially linking exposure location with downstream infection outcomes may prove logistically and analytically challenging (but not impossible, as shown by the MV Hondius outbreak; Redding et al., in review).

Furthermore, attempts to connect fine-scale behavioural exposure data with individual health outcomes can raise substantial ethical, privacy, and governance considerations, particularly under data protection frameworks such as GDPR [67,68]. These constraints may limit the feasibility of validating behavioural surveillance pipelines against individual-level transmission outcomes, at least without carefully designed participatory or public health collaborations. An important intermediate validation step may therefore be to test whether behavioural surveillance models can predict patterns of wildlife-human interactions themselves, rather than downstream disease outcomes. For example, models trained on existing data could identify locations predicted to exhibit high rates of aggression or other risky behaviours, with subsequent online data collection or targeted field observations used to assess whether these interactions emerge as expected. Our macaque case study is particularly well suited to this approach because several long-term field sites already collect detailed behavioural data on habituated populations (e.g. [75]). These systems provide an opportunity to test digital predictions by coordinating targeted behavioural observations at locations identified as high risk, thereby linking online behavioural signals with independent field-based measures of realised wildlife-human interactions. Validation like this would provide a direct test of

behavioural prediction while avoiding many of the logistical, ethical, and epidemiological challenges associated with linking exposure to individual infection outcomes.

Multilingual and translation issues may also arise [76], and an equitable approach to digital spillover observation will have to ensure that a range of languages are treated with respect and attention. We used autotranslated versions from google, with a wide vocabulary to detect any monkeys; this might be less viable in scenarios with more and more specific species identities, or where different species' names are used in place of each other in different languages. For example, rhesus macaques may be described as "monkeys" in English-language reviews, "bandar" in Hindi, or "monyet" in Indonesian, while related primates may be grouped under overlapping colloquial terms such as "kera" depending on local linguistic context. Such variation complicates automated species identification and may introduce substantial uncertainty into multilingual behavioural classification pipelines. This is particularly important where a given disease is associated primarily with one reservoir species rather than its close relatives: for example, reviews that identify "rodents" but do not specifically differentiate *Mastomys natalensis* will be less able to predict spillover rates of Lassa fever virus [77] unless the risky encounter rate is similar for *M. natalensis* and other rodents (which may not be true). The same is true for Herpes B virus in macaques versus sympatric langurs [78] or with Nipah virus in *Pteropus* fruit bats versus insectivorous bats [79].

We may also struggle to move between layers and granularity of prediction [45,80]: for example, identifying seasonally fluctuating interaction rates is only a first step to understanding outbreak seasonality [81]. It may be much harder to identify the genres of tourism associated with certain interactions at sub-seasonal scales and therefore use them to predict risk in the absence of this overall seasonal variation.

As with many digital data sources, it is necessary to handle the data and questions with sensitivity. It is very possible that using reviews of given tourist locations may risk angering people (and therefore potential backlash). Implications of "dirtiness" may risk influencing commerce and the economy via reputation, and it will be necessary to carry out the analyses and reports in an inequitable way to avoid vilifying people, reinforcing stereotypes, or being insensitive to cultural mores

[67,68]. In designing them, scientists likewise bear some responsibility to do so in a way that avoids possible use by bad actors for malicious purposes such as nonconsensual surveillance of individuals.

Moving forward: integrating behavioural traces into spillover risk prediction

The framework we propose represents only an initial step towards operationalising behavioural surveillance. The greatest value of these approaches will emerge when integrated alongside existing surveillance of pathogens, hosts, vectors, and environments rather than being treated as a standalone data stream. Several priorities are likely to accelerate this transition.

Quantifying the drivers of behavioural risk: Our case study demonstrates that behavioural risk can be detected, but not yet why it varies. An immediate priority will therefore be identifying the ecological, environmental, and anthropogenic drivers of risky wildlife-human interactions. For example, are rates of macaque aggression primarily explained by tourism intensity, provisioning, habitat fragmentation, population density, seasonality, or combinations of these factors? Such analyses would help distinguish the mechanisms generating behavioural hotspots and identify intervention points for reducing exposure before spillover occurs.

Evaluating the added value of behavioural information: Many existing spillover models already incorporate some element of host distributions, abundance, land use, climate, or human population density [17,27,55,57]. Fine-scale behavioural interaction rates are nested within these: for example, higher population density will correlate with higher encounter and interaction rates for many diseases [82]. Behavioural surveillance should therefore not be assumed to improve prediction automatically. Instead, an important research priority should be to determine when behavioural information provides predictive power beyond these existing ecological proxies. In some systems, fine-scale behavioural information may explain substantial additional variation in realised contact rates, whereas in others, local host abundance or environmental suitability may already capture much of the underlying risk; in these scenarios, digital surveillance of host abundances may nevertheless contribute something to our understanding (e.g., in rodent population booms [83,84]).

Expanding behavioural surveillance across taxa and systems: Our macaque case study represents one of the most conspicuous wildlife-human interfaces. Extending these approaches to additional systems, using a series of selected case studies, will determine their generality. Rabies, for example, represents a particularly attractive proof-of-concept pathogen because it has high spillover rates and the behavioural signatures (bites, aggression, stray dog encounters) are closely aligned with transmission risk [85,86]. Separate behavioural surveillance pipelines could potentially be developed for free-ranging dogs and wildlife reservoirs such as raccoons [87], allowing comparison of how different host species generate distinct patterns of human exposure risk. Other obvious opportunities include urban synanthropic species such as gulls, foxes, and pigeons, as well as wildlife tourism involving marine mammals, seabirds, and terrestrial megafauna. Different host-pathogen systems will probably require different digital signatures, with some relying (for example) primarily on direct interactions, while others rely on indirect indicators of environmental exposure.

Towards proactive and multimodal surveillance: Clearly, the end goal of behavioural surveillance should be not to rely solely on retrospective analysis of online text. Advances in natural language processing, computer vision, and multimodal machine learning may enable simultaneous extraction of behavioural information from text, photographs, and videos across multiple online platforms. Combined with prospective reporting tools or citizen science initiatives designed specifically to record wildlife-human interactions, these approaches could generate near-real-time behavioural surveillance systems to identify changes in exposure risk before outbreaks occur. Such steps would need to be carried out with significant ethical oversight to avoid misrepresentation or unbalanced impacts across social groups, avoiding mission creep and preventing their use for malicious purposes by bad actors.

Integrating behavioural surveillance with existing One Health infrastructure: Ultimately, behavioural surveillance should complement, rather than replace, existing surveillance approaches. Integrating behavioural indicators with pathogen surveillance, serology, wildlife monitoring, environmental data, and public health reporting offers the opportunity to develop a more mechanistic understanding of

spillover. Such integration would also provide a framework for validating behavioural predictions and identifying the ecological and social processes that translate wildlife-human interactions into pathogen transmission.

Concluding remarks

Behaviour is often taken for granted as a vital mechanism underlying emerging infectious disease risk [4], yet it remains one of the least systematically monitored components of One Health surveillance [21]. By showing that behavioural interactions can be recovered at scale from digital traces, we argue that behaviour itself can – and should – become a routine target of surveillance across the animal-human interface.

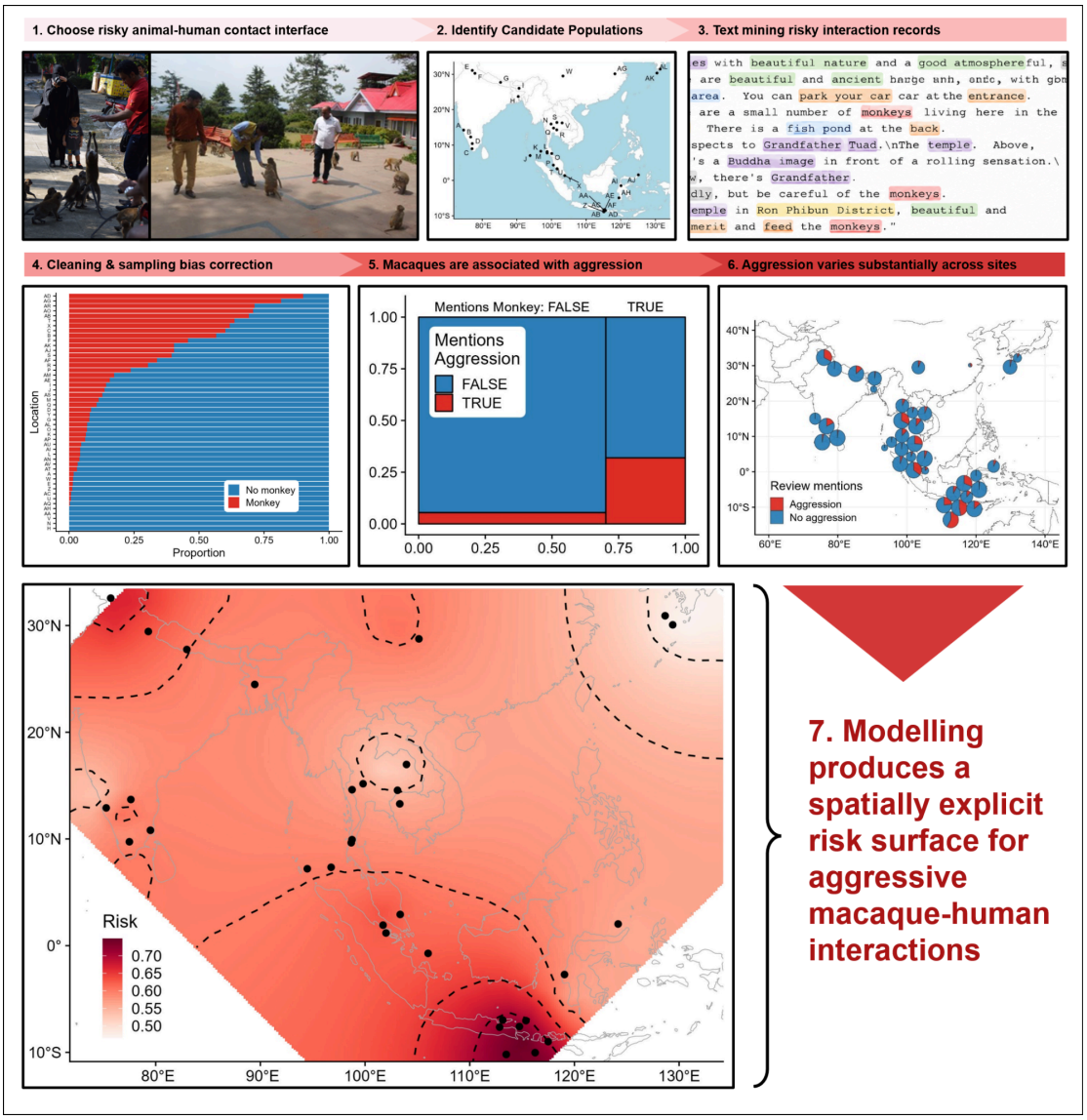

Figure: A case study demonstrating that behavioural risk for zoonotic spillover can be monitored using public digital posts. (1) Behavioural interfaces associated with transmission risk were first identified as a case study. Here, the example used were macaque monkeys, which are known to be present at a diversity of tourist locations. Photo credit: Krishna Balasubramaniam. (2) Candidate wildlife populations and human-animal interface sites were identified spatially. (3) Public online content, including reviews and social media posts, provides records of risky interactions and associated outcomes such as bites or scratches. (4) Natural language processing and text-mining approaches extract behavioural signals from large-scale unstructured text datasets. (5) Data cleaning and bias-correction procedures are applied to account for uneven reporting effort and sampling intensity. (6) Aggregated behavioural signals reveal population-level spatial patterns in risky interactions, allowing identification of behavioural hotspots potentially relevant to zoonotic spillover surveillance. (7) Putting these values through a spatiotemporally explicit INLA model produced a two-dimensional risk surface with substantial variation in risk. Points represent sampled populations (with jitter added); deeper red colours denote higher predicted risk.

References

1. Jones, K.E. *et al.* (2008) Global trends in emerging infectious diseases. *Nature* 451, 990–993
2. Morse, S.S. *et al.* (2012) Prediction and prevention of the next pandemic zoonosis. *Lancet* 380, 1956–1965
3. Woolhouse, M.E.J. *et al.* (2005) Emerging pathogens: The epidemiology and evolution of species jumps. *Trends Ecol. Evol.* 20, 238–244
4. Plowright, R.K. *et al.* (2017) Pathways to zoonotic spillover. *Nature Publishing Group*
5. Fagre, A.C. *et al.* (2022) Assessing the risk of human-to-wildlife pathogen transmission for conservation and public health. *Ecol. Lett.* 25, 1534–1549
6. Van Hamme, G. *et al.* (2021) Keep your distance: Using Instagram posts to evaluate the risk of anthroponotic disease transmission in gorilla ecotourism. *People Nat. (Hoboken)* 3, 325–334
7. Morgan, M. and Palmer, J. (2007) Dog bites. *BMJ* 334, 413–417
8. Engel, G.A. *et al.* (2002) Human exposure to herpesvirus B-seropositive macaques, Bali, Indonesia. *Emerg. Infect. Dis.* 8, 789–795
9. Admassu, B. *et al.* (2015) Review on bovine tuberculosis. *Eur. J. Biol. Sci.* 7, 169–185
10. Talmon, I. *et al.* (2025) Using wild-animal tracking for detecting and managing disease outbreaks. *Trends Ecol. Evol.* 40, 760–771
11. McLean, H.E. *et al.* (2022) Social media as a window into human-wildlife interactions and zoonotic disease risk: an examination of wild pig hunting videos on YouTube. *Hum. Dimensions Wildl.* 27, 307–320
12. Carlson, C.J. *et al.* (2021) The future of zoonotic risk prediction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 376
13. Grange, Z.L. *et al.* (2021) Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2002324118
14. Plourde, B.T. *et al.* (2017) Are disease reservoirs special? Taxonomic and life history characteristics. *PLoS One* 12, e0180716
15. Albery, G.F. *et al.* (2020) Predicting the global mammalian viral sharing network using phylogeography. *Nat. Commun.*
16. Forero-Muñoz, N.R. *et al.* (2024) The coevolutionary mosaic of bat betacoronavirus emergence risk. *Virus Evol* 10, vead079
17. Olival, K.J. *et al.* (2017) Host and viral traits predict zoonotic spillover from mammals. *Nature* 546, 646–650
18. Mollentze, N. and Streicker, D.G. (2023) Predicting zoonotic potential of viruses: where are we? *Curr. Opin. Virol.* 61, 101346
19. Betke, B. *et al.* (2023) Bat anthropogenic roosting ecology influences taxonomic and geographic predictions of zoonotic risk*bioRxiv*, bioRxiv, 2023.12. 12.571362
20. Albery, G.F. *et al.* (2022) Urban-adapted mammal species have more known pathogens. *Nat. Ecol. Evol.* 6, 794–801
21. Ahmed, A.N. *et al.* (2025) Human animal contact, land use change and zoonotic disease risk: a protocol for systematic review. *Syst. Rev.* 14, 65
22. Luby, S.P. *et al.* (2009) Transmission of human infection with Nipah virus. *Clin. Infect. Dis.* 49, 1743–1748
23. Forbes, K.M. *et al.* (2018) Hantavirus maintenance and transmission in reservoir host populations. *Curr. Opin. Virol.* 28, 1–6
24. Mann, E. *et al.* (2015) A review of the role of food and the food system in the transmission and spread of ebolavirus. *PLoS Negl. Trop. Dis.* 9, e0004160
25. Bansal, S. *et al.* (2007) When individual behaviour matters: homogeneous and network models in epidemiology. *J. R. Soc. Interface* 4, 879–891
26. Firth, J.A. *et al.* (2020) Using a real-world network to model localized COVID-19 control strategies. *Nat. Med.* 26, 1616–1622

27. Carlson, C.J. *et al.* (2022) Climate change increases cross-species viral transmission risk. *Nature* DOI: 10.1101/2020.01.24.918755
28. Jarić, I. *et al.* (2020) IEcology: Harnessing large online resources to generate ecological insights. *Trends Ecol. Evol.* 35, 630–639
29. Mikula, P. *et al.* (2026) Harnessing iEcology data to uncover invasive species behaviour. *Methods Ecol. Evol.* 17, 1022–1031
30. Salathé, M. *et al.* (2012) Digital epidemiology. *PLoS Comput. Biol.* 8, e1002616
31. Eysenbach, G. (2009) Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J. Med. Internet Res.* 11, e11
32. Ginsberg, J. *et al.* (2009) Detecting influenza epidemics using search engine query data. *Nature* 457, 1012–1014
33. Hausmann, A. *et al.* (2018) Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conserv. Lett.* 11, e12343
34. Hertzog, L.R. *et al.* (2021) Model-based integration of citizen science data from disparate sources increases the precision of bird population trends. *Divers. Distrib.* 27, 1106–1119
35. Toivonen, T. *et al.* (2019) Social media data for conservation science: A methodological overview. *Biol. Conserv.* 233, 298–315
36. Wood, S.A. *et al.* (2013) Using social media to quantify nature-based tourism and recreation. *Sci. Rep.* 3, 2976
37. Di Minin, E. *et al.* (2015) Prospects and challenges for social media data in conservation science. *Front. Environ. Sci.* 3
38. Harrington, L.A. *et al.* (2026) The role of social media in promoting inappropriate wildlife tourism: a case study of direct contact animal-visitor interactions portrayed on YouTube. *J. Sustain. Tour.*
39. Farrell, M.J. *et al.* (2022) Past and future uses of text mining in ecology and evolution. *Proc. Biol. Sci.* 289, 20212721
40. Supp, S.R. *et al.* (2021) Estimating the movements of terrestrial animal populations using broad-scale occurrence data. *Mov. Ecol.* 9, 60
41. Mason, B.M. *et al.* (2025) iNaturalist accelerates biodiversity research. *Bioscience* 75, 953–965
42. Skotis, A. *et al.* (2023) Identifying drivers of evaluation bias in online reviews of city destinations. *International Journal of Information Management Data Insights* 3, 100184
43. Correia, R.A. *et al.* (2021) Digital data sources and methods for conservation culturomics. *Conserv. Biol.* 35, 398–411
44. Plowright, R.K. *et al.* (2021) Land use-induced spillover: a call to action to safeguard environmental, animal, and human health. *The Lancet Planetary Health* 5, e237–e245
45. Becker, D.J. *et al.* (2019) The problem of scale in the prediction and management of pathogen spillover. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 374, 20190224
46. Sims, N. and Kasprzyk-Hordern, B. (2020) Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level. *Environ. Int.* 139, 105689
47. Kirby, A.E. *et al.* (2021) Using wastewater surveillance data to support the COVID-19 response - United States, 2020-2021. *MMWR Morb. Mortal. Wkly. Rep.* 70, 1242–1244
48. Beauchamp, N. (2022) "this candle has no smell": Detecting the effect of COVID anosmia on Amazon reviews using Bayesian vector autoregression. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1363–1367
49. Priston, N.E.C. and McLennan, M.R. (2013) Managing humans, managing macaques: Human–macaque conflict in Asia and Africa. In *The Macaque Connection*, pp. 225–250, Springer New York
50. Anthony, S.J. *et al.* (2015) Non-random patterns in viral diversity. *Nat. Commun.* 6, 1–7
51. De Moor, D. *et al.* (2025) MacaqueNet: Advancing comparative behavioural research through large-scale collaboration. *J. Anim. Ecol.* 94, 519–534
52. Ratcliffe, R. (2022) With the tourists gone, what now for Lopburi's famous monkeys? *The*

Guardian, The Guardian
53. Di Cecco, G.J. *et al.* (2021) Observing the observers: How participants contribute data to iNaturalist and implications for biodiversity science. *Bioscience* 71, 1179–1188
54. Hu, N. *et al.* (2009) Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 144–147
55. Gibb, R. *et al.* (2020) Zoonotic host diversity increases in human-dominated ecosystems. *Nature* DOI: 10.1038/s41586-020-2562-8
56. Karesh, W.B. *et al.* (2012) Ecology of zoonoses: natural and unnatural histories. *Lancet* 380, 1936–1945
57. Allen, T. *et al.* (2017) Global hotspots and correlates of emerging zoonotic diseases. *Nat. Commun.* 8, 1124
58. Maréchal, L. *et al.* (2016) Primates' behavioural responses to tourists: evidence for a trade-off between potential risks and benefits. *Sci. Rep.* 6, 32465
59. Cui, Q. *et al.* (2021) The escalating effects of wildlife tourism on human-wildlife conflict. *Animals (Basel)* 11, 1378
60. Leca, J.-B. *et al.* (2021) Acquisition of object-robbing and object/food-bartering behaviours: a culturally maintained token economy in free-ranging long-tailed macaques. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 376, 20190677
61. Guenther, S.K. and Shanahan, E.A. (2020) Communicating risk in human-wildlife interactions: How stories and images move minds. *PLoS One* 15, e0244440
62. Usui, R. *et al.* (2014) Park rangers' behaviors and their effects on tourists and Tibetan macaques (Macaca thibetana) at Mt. Huangshan, China. *Animals (Basel)* 4, 546–561
63. Lazer, D. *et al.* (2014) Big data. The parable of Google Flu: traps in big data analysis. *Science* 343, 1203–1205
64. Palmer, J.R.B. *et al.* (2017) Citizen science provides a reliable and scalable tool to track disease-carrying mosquitoes. *Nat. Commun.* 8, 916
65. Kianfar, N. *et al.* (2026) The role of citizen science in mosquito-borne disease surveillance and control: A scoping review. *PLoS One* 21, e0348697
66. Simoncini, A. *et al.* (2025) Taxonomic bias towards charismatic and easy-to-find mammals shapes knowledge of parasites. *Biodivers. Conserv.* DOI: 10.1007/s10531-025-03147-1
67. Mittelstadt, B.D. *et al.* (2016) The ethics of algorithms: Mapping the debate. *Big Data Soc.* 3, 205395171667967
68. Floridi, L. and Taddeo, M. (2016) What is data ethics? *Philos. Trans. A Math. Phys. Eng. Sci.* 374, 20160360
69. Jonsson, C.B. *et al.* (2010) A global perspective on hantavirus ecology, epidemiology, and disease. *Clin. Microbiol. Rev.* 23, 412–441
70. Vial, P.A. *et al.* (2023) Hantavirus in humans: a review of clinical aspects and management. *Lancet Infect. Dis.* 23, e371–e382
71. Luby, S.P. *et al.* (2006) Foodborne Transmission of Nipah Virus, Bangladesh - Volume 12, Number 12—December 2006 - Emerging Infectious Diseases journal - CDC. *Emerg. Infect. Dis.* 12, 1888–1894
72. Islam, M.S. *et al.* (2016) Nipah virus transmission from bats to humans associated with drinking traditional liquor made from date palm sap, Bangladesh, 2011-2014. *Emerg. Infect. Dis.* 22, 664–670
73. Tatem, A.J. *et al.* (2006) Global transport networks and infectious disease spread. *Adv. Parasitol.* 62, 293–343
74. Brockmann, D. and Helbing, D. (2013) The hidden geometry of complex, network-driven contagion phenomena. *Science* 342, 1337–1342
75. Balasubramaniam, K.N. *et al.* (2020) Impact of individual demographic and social factors on human-wildlife interactions: a comparative study of three macaque species. *Sci. Rep.* 10, 21991
76. Hedderich, M.A. *et al.* (2021) A survey on recent approaches for natural language processing in low-resource scenarios. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Online
77. Garry, R.F. (2023) Lassa fever - the road ahead. *Nat. Rev. Microbiol.* 21, 87–96
78. Hilliard, J. (2007) Monkey B virus. *Human herpesviruses: biology, therapy, and immunoprophylaxis* at <https://www.ncbi.nlm.nih.gov/books/NBK47426/>
79. Lee, B. and Rota, P.A., eds. (2012) *Henipavirus: Ecology, molecular virology, and pathogenesis*, ((2012th edn)), Springer
80. Levin, S.A. (1995) The problem of pattern and scale in ecology. In *Ecological Time Series*, pp. 277–326, Springer US
81. Altizer, S. *et al.* (2006) Seasonality and the dynamics of infectious diseases: Seasonality and infectious diseases. *Ecol. Lett.* 9, 467–484
82. Albery, G.F. *et al.* (2025) Density-dependent network structuring within and across wild animal systems. *Nat. Ecol. Evol.* DOI: 10.1038/s41559-025-02843-z
83. Friant, S. *et al.* (2025) Reducing the threats of rodent-borne zoonoses requires an understanding and leveraging of three key pillars: disease ecology, synanthropy, and rodentation. *Lancet Planet. Health*
84. Ecke, F. *et al.* (2022) Population fluctuations and synanthropy explain transmission risk in rodent-borne zoonoses. *Nat. Commun.* 13, 7532
85. Hampson, K. *et al.* (2015) Estimating the Global Burden of Endemic Canine Rabies. *PLoS Negl. Trop. Dis.* 9, 1–20
86. Mancy, R. *et al.* (2022) Rabies shows how scale of transmission can enable acute infections to persist at low prevalence. *Science* 376, 512–516
87. Slate, D. *et al.* (2008) Attaining raccoon rabies management goals: history and challenges. *Dev. Biol. (Basel)* 131, 439–447