Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mapping digital disease surveillance research in Africa using bibliometric analysis and structural topic modelling

Cornelius Ayantse

cornelius.ayantse.25@ucl.ac.uk

University College London

Rotimi F. Afolabi

University of Ibadan

OlaOluwa S. Yaya

University of Ibadan

Deborah O. Oluwasola

McPherson University

Systematic Review

Keywords: digital disease surveillance, Africa, bibliometric analysis, structural topic modelling, health information systems, research equity

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11164948/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Mapping digital disease surveillance research in Africa using bibliometric analysis and structural topic modelling

Cornelius Ayantse
Institute for Global Health, University College London, London, United Kingdom and Department of Health Planning, Research and Statistics, Benue State Ministry of Health and Human Services, Benue, Nigeria
Email address: cornelius.ayantse.25@ucl.ac.uk

Rotimi F. Afolabi
Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan, Ibadan, Nigeria
Email address: rotimifelix@yahoo.com

OlaOluwa S. Yaya
Computational Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria & Centre for Econometrics and Applied Research, Ibadan, Nigeria
Email address: os.yaya@ui.edu.ng

Deborah O. Oluwasola
Public Health Unit, Faculty of Nursing Science, McPherson University, Ogun State, Nigeria
Email address: oluwasolaolusola@yahoo.com

Abstract

Introduction. Surveillance in Africa has transited from paper registers under the Integrated Disease Surveillance and Response strategy to electronic case reporting, mobile and community reporting, geospatial analysis and genomic pipelines.
Methods. We therefore searched Scopus on 25 July 2026, screened 600 publications against stated criteria to 563 documents, and analysed them by bibliometric mapping and six-topic structural topic models, with prevalence estimated by publication year. Each document was coded for the African countries it studied, the diseases it addressed, and the technologies it used. The attention each country received was compared with the presence of authors' affiliations. The research spanned three decades, dispersed across four disciplines that were not mapped.
Results. Publication outputs grew at 14.9% annually with no detectable structural break, even during the pandemic periods (Ebola and COVID-19). Sources rose from 38 to 167, and countries increased from 20 to 46. About 26.5% were cited within the corpus, and four of the six most co-cited references were institutional documents rather than peer-reviewed research. Digital health innovation increased marginally annually, while clinical HIV data declined. Geospatial analysis and machine learning dominated the studies. Local authorship is prominent in the studies.
Conclusion. This field of Digital Disease Surveillance has grown steadily in Africa, and it consists of short, unconnected studies citing policy instruments rather than addressing analytical methods or interoperability implementation reviews. Continuity in the subject matter is therefore recommended.

Keywords: digital disease surveillance; Africa; bibliometric analysis; structural topic modelling; health information systems; research equity

1. Introduction

The speed at which an outbreak is detected and reported determines how far it spreads. Africa is prone to this as the continent carries a heavy burden of epidemic-prone diseases. The World Health Organisation (WHO) African Region adopted the Integrated Disease Surveillance and Response (IDSR) strategy in 1998 to strengthen disease detection, reporting, analysis and response (Regional Committee for Africa, 1998; Perry et al., 2007). During the period, alerts were logged in paper registers and sent via hierarchical reporting chains, and disease-specific programs typically kept separate reporting systems. According to reports from Ghana, Malawi, Nigeria and Zambia, these arrangements led to delays in disease reporting and response (Joseph et al., 2018; Adokiya et al., 2015; Mandyata et al., 2017; Ibrahim et al., 2020). Since then, improvements to digital platforms over the next 20 years have included the District Health Information Software (DHIS2) and national adaptations, electronic case-based reporting via mobile devices, and laboratory information systems that are increasingly linked to genomic sequencing platforms (Mustafa et al., 2023; Kaburi et al., 2024). These have led to significant institutional and technological shifts in disease surveillance in Africa, from disjointed paper-based systems to increasingly integrated and digitalised surveillance architectures.

The use of digital surveillance technologies has since increased due to major pandemics. Mobile applications for case monitoring, contact tracing, and outbreak reporting were developed in response to the 2014–2016 West African Ebola epidemic (Tom-Aba et al., 2018). Continental disease surveillance and outbreak preparedness were further reinforced in 2017 with the creation of the Africa Centres for Disease Control and Prevention (Africa CDC) (African Union, 2017). In addition to significant investments in laboratory and data infrastructures, the COVID-19 pandemic led to an unprecedented growth of digital contact tracing, genomic surveillance, and electronic reporting throughout Africa (Tegally et al., 2022). More recently, in August 2024, the Africa Centres for Disease Control and Prevention declared mpox a Public Health Emergency of Continental Security, highlighting ongoing shortcomings in contact tracing, testing, and surveillance capabilities and reiterating the strategic significance of digital surveillance systems (Africa Centres for Disease Control and Prevention, 2024).

As a result, the scope of digital disease surveillance has widened considerably, as it includes mobile and community-based reporting, laboratory and genomic surveillance, geospatial and remote-sensing techniques, electronic health information and reporting systems, and computational techniques like machine learning and other artificial intelligence (Aku et al., 2023; Silenou et al., 2021; Phoobane et al., 2022).

Current evidence syntheses offer insightful but disjointed descriptions of this field of study. In their systematic review of published evidence on the digitalisation of infectious disease surveillance in Africa between 2003 and 2022, Kaburi et al. (2024) found only 29 eligible reports describing 27 projects across 13 countries, about 75% of which were pilot projects. Certain countries (Mustafa et al., 2023), particular technology classes (Aku et al., 2023), or specific surveillance functions and methodological techniques (Silenou et al., 2021) have all been the subject of other research. The production and organisation of the related scientific literature have not received as much attention as surveillance performance and implementation difficulties in reviews of IDSR implementation (Wolfe et al., 2021; Mremi et al., 2021). While bibliometric studies have looked at related topics, such as digital health research in a few African nations (Agbeyangi et al., 2026) and infectious disease prediction in Africa (Phoobane et al., 2022), they do not specifically look at digital disease surveillance as a public health function. No study has therefore mapped the research literature on digital disease surveillance in Africa across the full period over which the region's surveillance systems were digitised.

Methodological limitations are another drawback of the current study, as prior bibliometric evaluations have mostly focused on descriptive indicators, citation analysis, co-occurrence networks, and author-generated keywords. These methods help map out obvious connections within a work of literature. Still, they might only give a partial picture of the underlying thematic structure, especially when keywords are absent or given inconsistently across disciplines. Examining the literature's latent structure directly from textual content is made possible by a probabilistic topic-modelling approach. Specifically, Structural Topic Modelling (STM) models topic prevalence as a function of document-level attributes such as publication year, while representing each document as a mixture of latent topics (Roberts et al., 2014; Roberts et al., 2019). Its application to digital disease surveillance research can therefore reveal how the prominence of those themes has changed over time.

Studies pertaining to a specific African nation do not always involve scholars connected to institutions in that nation, and research carried out in Africa is not always headed by researchers headquartered in African institutions. This distinction is crucial because there have long been significant disparities in research capacity and participation in African health research due to patterns of international research collaboration and authorship (Hedt-Gauthier et al., 2019; Mbaye et al., 2019; Kisenge et al., 2021; Ayantse et al., 2024). As a result, this study makes a distinction between the nation of author affiliation, which refers to the

institutional location of the scholars contributing to the publication, and the country of study, which refers to the African country or countries covered in a publication.

Thus, by integrating bibliometric analysis, scientific mapping, and structural topic modelling, the current work analyses 563 documents published between 1993 and 2025 and analyses the evolution and intellectual structure of digital disease surveillance research in Africa. In order to supplement implementation-oriented reviews with an analysis of how knowledge in this subject has been produced, related, and thematically arranged, the study takes the literature itself as the focus of analysis (Cobo et al., 2011; Aria and Cuccurullo, 2017; Donthu et al., 2021). The study specifically answers the following six questions: (1) Is there a discernible shift in the growth trajectory of research output on digital disease surveillance in Africa between 1993 and 2025? (2) Which authors, countries, institutions, and publication sources have made the most contributions to the topic, and how do they collaborate? (3) What research areas and theoretical underpinnings support the literature? (4) What hidden themes are present in research on digital disease surveillance, how common are they, and how have they changed over time? (5) Which African countries, diseases and digital technologies are discussed in the literature, and how has the emphasis on technology evolved over time? (6) How much does the presence of researchers connected to institutions in certain African nations correlate with the amount of research focus given to such nations?

Three main contributions are made by the study. First, it offers a thorough bibliometric description of a field of study that has historically been mostly investigated through reviews that are country-, technology-, disease-, or implementation-specific. Second, by using Structural Topic Modelling to find latent topic patterns and track their development over time, it expands on traditional science mapping. Third, it offers a more nuanced evaluation of the global distribution of research interest and local research engagement by separating countries of study from countries of author affiliation. When taken as a whole, these contributions offer a methodical evidence base for comprehending the growth of digital disease surveillance research in Africa as well as the field's emerging directions, knowledge structures, thematic goals, and geographic disparities.

2. Materials and Methods

2.1 Data, PRISMA and Inclusions

Scopus database was searched, and the query below was run and returned 600 records. The search combined three concept blocks with AND: digital, electronic and computational methods; the surveillance function, as explicit phrases; and African geography, comprising the continental term together with fourteen individually named countries. It was applied to title, abstract and keyword fields and restricted to journal articles and reviews in English, published from 1991 to 2025.

```
TITLE-ABS-KEY ( ( "digital" OR "digital health" OR "ehealth" OR "e-health" OR
"mhealth" OR "m-health" OR "digital surveillance" OR "digital disease surveillance"
OR "electronic surveillance" OR "web-based" OR "internet-based" OR interoperab*
OR "health information exchange" OR "information system*" OR "health information
system*" OR "electronic health record*" OR "data platform*" OR "data integration"
OR "data sharing" OR "artificial intelligence" OR "machine learning" OR "big data"
OR "geographic information system*" OR DHIS2 ) AND ( "disease surveillance" OR
"public health surveillance" OR "epidemiological surveillance" OR "syndromic
surveillance" OR "event-based surveillance" OR "integrated disease surveillance" OR
"outbreak detection" OR "disease monitoring" OR "epidemic intelligence" )
AND ( africa* OR "sub-saharan africa*" OR nigeria* OR kenya* OR ghana* OR
ethiopia* OR uganda* OR tanzania* OR "south africa*" OR rwanda* OR senegal* OR
cameroon* OR zambia* OR malawi* OR mozambique* OR "democratic republic of
congo" ) ) AND PUBYEAR > 1990 AND PUBYEAR < 2026 AND ( LIMIT-TO (
SRCTYPE , "j" ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE ,
"re" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

Figure 1 gives the selection process. Of 600 records retrieved, one was in press and one a duplicate, leaving 598 that met the inclusion criteria; a further 28 were excluded under the exclusion criteria in Table 1, giving 570 eligible documents. Seven carry no usable abstract and were removed, so the analyses reported here use the 563 abstract-bearing documents throughout. Table 2 gives the resulting screening flow.

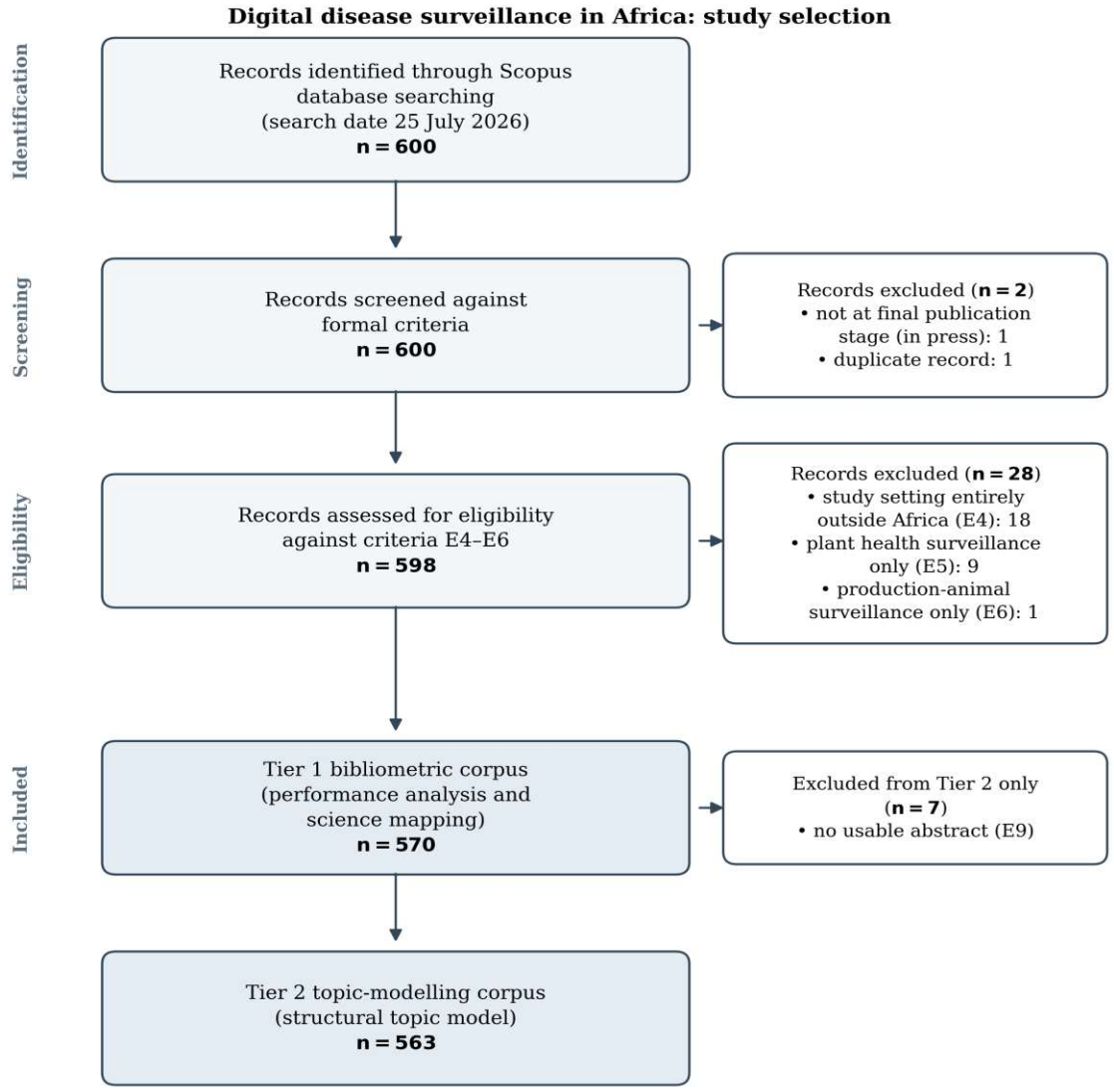


Figure 1. PRISMA flow diagram of the study selection process

Table 1. Inclusion criteria and, applied to the records meeting them, exclusion criteria

	Criterion	**Operationalisation**
Inclusion criteria		
I1	Satisfies the geographic, surveillance and digital conditions simultaneously	Title, abstract and keyword screening
I2	Peer-reviewed journal article or review	Document Type
I3	Published in English	Language of Original Document
I4	Published 1993 to 2025 inclusive	Year
I5	Final publication stage	Publication Stage
I6	Carries a usable English abstract	Abstract of 100 characters or more, excluding placeholder text

	Criterion	Operationalisation
Exclusion criteria, applied to records meeting I1 to I6		
E1	Study setting entirely outside Africa	Fails the geographic condition, including surveillance of African-origin cases conducted elsewhere
E2	Surveillance target exclusively plant or crop disease	Fails the surveillance condition: no human health outcome
E3	Non-zoonotic production-animal disease with no human health or One Health framing	Fails the surveillance condition: agricultural rather than public health surveillance
E4	Duplicate record	One record per publication; the copy with the longer cited-reference field was retained

Table 2. Screening flow from records identified to documents analysed

Stage	Records	Note
Records identified	600	Single Scopus query
Did not meet inclusion criteria	2	One article in press (I5); no record failed I2, I3 or I4, since the query applied those limits
E1 setting outside Africa	18	Includes imported malaria surveillance in China, tuberculosis institute activity in Korea, SARS-CoV-2 variant surveillance in Brazil
E2 plant health only	9	Includes banana disease detection from aerial imagery and unmanned aerial systems for crop disease
E3 production animal only	1	Of 33 records mentioning a non-zoonotic production-animal pathogen, 32 were framed within One Health platforms and retained.
E4 duplicate record	1	Same DOI indexed twice, as 2023 and 2024
Eligible documents	570	Retained after exclusion criteria applied
No usable abstract (I6)	7	Editorials, consensus statements and profiles indexed without an abstract
Documents analysed	563	Abstract-bearing corpus used throughout

Records identified 600; excluded at formal screening 2; excluded under the eligibility rules 28; excluded for want of an abstract 7. The 28 records excluded under E4 to E6 are available from the corresponding author, listed individually with the criterion applied to each.

Identification, screening and eligibility were reported using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow (Page et al., 2021), adapted for bibliometrics rather than an effectiveness review and used solely to make the path from records retrieved to records analysed auditable. No risk-of-bias assessment or effect synthesis was

undertaken. Records were drawn from the Scopus database, searched once on 25 July 2026.[1] The analyses reported included: document co-citation, bibliographic coupling, reference publication year spectroscopy and the Price index.[2] In addition, the English-language restriction excluded the French and Portuguese literature, although 173 of the retained records reported studies conducted in francophone African countries, so those settings were represented.

Further from the screening in Tables 1 and 2, within the digital block, three terms occurred prominently: “information system” appeared in 283 of 563 documents, “digital” in 124 and “geographic information system” in 87, while “data platform” appeared in two and “health information exchange” in four. Within the geographic block, the continental terms retrieved 378 documents.

A record was eligible if it reported, developed, evaluated or reviewed digital, electronic or computational data, tools or systems for the detection, notification, monitoring, verification, analysis or forecasting of human disease or public health events in one or more African countries. Eligibility required three conditions to hold simultaneously: geographic, surveillance and digital; records satisfying only two did not meet the inclusion criteria. Table 1 sets out the criteria in full. Paper-based surveillance qualified only where the contribution concerned the transition to digital reporting.

Health and zoonotic surveillance was eligible where the pathogen is zoonotic, or the platform explicitly serves human health protection. Of 33 records mentioning a non-zoonotic production-animal pathogen, 32 were retained on this basis. Plant health surveillance was not eligible, excluding nine records. Geography was judged on the study setting rather than authorship, excluding eighteen records reporting surveillance conducted wholly in another region. Computational methods counted as digital whether or not the word appeared, so geospatial, forecasting and machine-learning studies were retained. Screening was carried out on titles, abstracts and keywords, with ambiguous records read in full and resolved by two authors.

[1] Scopus indexes a wider set of journals than Web of Science, with better coverage of health and of regionally published research, and it carries the structured affiliation and cited-reference fields on which the country-level and reference-based analyses depend (Baas et al., 2020; Mongeon et al., 2016). Although web of science does supply references, it formats the reference string differently from Scopus; so, merging the two fragments a single cited work into two entities and degrades precisely the analyses that depend on it.

[2] However, database such as PubMed does not supply cited-reference data in a form that bibliometric software can read; thus, adding it would extend none of the four analyses.

2.2 Corpus construction

Two data-handling decisions are recorded. Where a publication appeared twice, the copy with the longer cited-reference field was retained, since co-citation depends entirely on that field. Source titles contained inconsistent whitespace that split individual journals in two; whitespace was normalised in 12 records, reducing distinct sources from 262 to 259. The resulting corpus spans 1993 to 2025 across 232 sources with 12,067 citations.

2.3 Bibliometric analysis

Indicators were computed with the bibliometrix package (Aria et al., 2017). Every measure named below is defined formally in Table 3. Author productivity is reported as a whole and in fractionalised counts and tested against Lotka's inverse-square law (Lotka, 1926); source concentration is assessed by Bradford's law (Bradford, 1934); and influence is summarised by the h-index (Hirsch, 2005). Country output is reported for corresponding-author countries with Single-Country Publication (SCP) and Multiple-Country Publication (MCP) counts. Citation impact is reported three ways: the global citation score from Scopus, the local citation score counting citations from within the corpus, which distinguishes documents that are influential in this field from those widely cited elsewhere, and citations per year since publication.

Table 3. Formal definition of every measure reported in the main text

Measure	Definition	Interpretation
Lotka's law	$f(x) = C / x^{\beta}$, where $f(x)$ is the proportion of authors with x publications	β near 2 is the classical inverse-square distribution of productivity
Lotka goodness of fit	$D = \max \lvert F_o(x) - F_t(x) \rvert$, compared with $1.63 / \sqrt{n}$	Pao's test; D above the critical value rejects the inverse-square law
Bradford's law	Sources ranked by output and partitioned into zones of equal yield with sizes $1 : n : n^2$	Larger n means output is more concentrated in a small core of journals
h-index	Largest h such that h documents each have at least h citations	Joint measure of output and impact
Multiple-country ratio	$MCP_i / (SCP_i + MCP_i)$ for country i	Share of a country's output that is internationally co-authored
Local citation score	LCS_i = citations to document i from within the corpus	Influence inside the field, as against global visibility
Age-normalised citations	$TC_i / (2025 - PY_i + 1)$	Citation rate per year of exposure
Co-citation strength	c_{ij} = number of documents citing references i and j together	Shared intellectual foundations
Bibliographic coupling	$b_{ij} = \lvert R_i \cap R_j \rvert$, the overlap of the reference lists of documents i and j	Proximity of current research activity
Association strength	$s_{ij} = c_{ij} / (c_i\, c_j)$, applied to every co-occurrence matrix before clustering	Removes the effect of raw frequency on link weight
Cluster centrality	$c = 10 \times \Sigma\, s_{kl}$, summed over links from cluster k to all other clusters l	External connectedness of a theme: its relevance to the field
Cluster density	$d = 100 \times (\Sigma\, s_{ij}$ within the cluster$) / w$, for w keywords in the cluster	Internal cohesion of a theme: its degree of development
Inclusion index	$I = \lvert K_t \cap K_{t+1} \rvert / \min(\lvert K_t \rvert, \lvert K_{t+1} \rvert)$	Strength of thematic continuity between adjacent periods
Price index	Share of a document's references published within the preceding five years	Reliance on recent as against classic literature
RPYS deviation	$n_y - \text{median}(n_{y-2} \ldots n_{y+2})$, for cited-reference counts by year y	Peaks locate the years in which foundational work appeared
Local authorship rate	Share of documents studying country c that include at least one author affiliated in c	Participation of local institutions in research about their own country
Exclusivity	Rank of $\beta_{kv} / \Sigma_j\, \beta_{jv}$ among the leading terms of topic k	How far a topic's leading terms are specific to it
FREX	Harmonic mean of the frequency and exclusivity ranks, weighted at $\omega = 0.5$	Selects terms that are both frequent in a topic and distinctive to it

Association strength follows van Eck and Waltman; centrality and density follow Callon and colleagues as implemented for thematic mapping; co-citation follows Small and bibliographic coupling follows Kessler; the local authorship rate adapts the affiliation-assignment approach of Kisenge and colleagues. Full citations are in the reference list of the main text.

Growth was tested rather than described. Let y_t be the number of documents published in year t. The logarithm of annual output was modelled as a piecewise linear function of time with m unknown break dates, so that within segment j

$$\log y_t = \alpha_j + \beta_j t + \varepsilon_t, \quad T_{j-1} < t \le T_j \tag{1}$$

and the break dates were estimated by the Bai-Perron dynamic programming algorithm (Bai and Perron, 2003) implemented in strucchange (Zeileis et al., 2002), with the supremum F statistic testing the null of no break. . Because that test scans all candidate dates and has limited power on a short annual series, level shifts and slope changes at the two dates the policy narrative would predict, 2014 and 2020, were also tested directly.The supremum F test scans all candidate dates and so has limited power on a short annual series. Level shifts and slope changes were therefore tested directly at 2014 and 2020, the two dates the policy narrative would predict

Five mapping analyses were carried out, all with co-occurrence matrices normalised by association strength and communities detected by the Louvain algorithm: a three-field plot linking authors, keywords and sources; keyword co-occurrence; a co-word thematic map placing clusters on centrality and density (Cobo et al., 2011; Callon et al., 1991), with thematic evolution across three periods bounded at the Ebola and COVID-19 pandemic dates; document co-citation for the intellectual base (Small, 1973) and bibliographic coupling for the research front (Kessler, 1963); and co-authorship networks at country and author level. Reference publication year spectroscopy (RPYS) (Marx et al., 2014) and the Price index (Price, 1970) were used to locate foundational years and measure reliance on recent as against classic work.

2.4 Structural topic model

Abstracts were modelled with a structural topic model (Roberts et al., 2014; Roberts et al., 2019). Let D be the number of documents, K the number of topics and X_e a row of the prevalence design matrix. Figure 2 gives the generative structure of the structural topic model.[3] For each document, a vector of $K - 1$ latent parameters is drawn from a multivariate normal whose mean depends on the covariates, and mapped to the simplex by the logistic transformation:

$$\eta_d \sim \text{Normal}_{K-1}(X_d\gamma, \Sigma), \quad \theta_{dk} = \exp(\eta_{dk}) / \Sigma_j \exp(\eta_{dj}) \tag{2}$$

[3] Shaded nodes are observed. X_e is the matrix of prevalence covariates, a spline basis in publication year; η_e is the logistic-normal parameter vector, θ_e the document-topic proportions, z the token-level topic assignment, w the observed word and β_k the topic-word distribution.

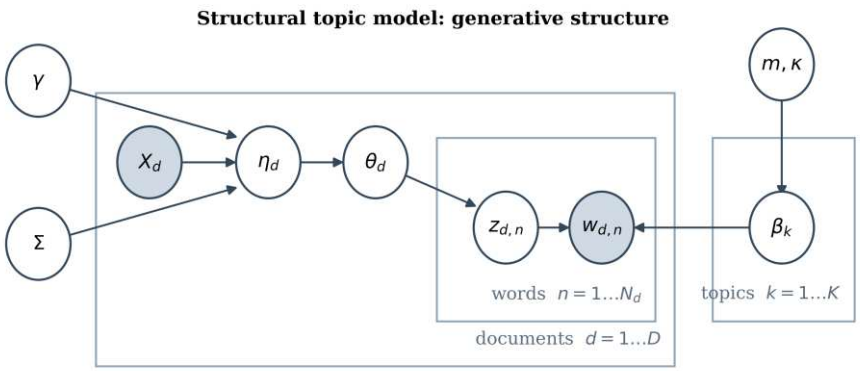


Figure 2. Graphical model of the structural topic

Word probabilities within each topic take the form $\beta_{kv} \propto \exp(m_v + \kappa_{kv})$, and each token is generated by drawing a topic from θ_e and then a word from that topic. Content covariates are available in the general model but are not used here. Setting X_e to a column of ones and replacing the logistic-normal prior with a Dirichlet distribution reduces the model to Latent Dirichlet Allocation (Blei et al., 2003). The two retained features are the reason for preferring the structural form, since the covariate matrix carries publication year and produces the estimated trajectories, and the unrestricted covariance Σ permits the topic correlation network to be estimated at all, a Dirichlet prior forcing topics to be very nearly independent (Blei and Lafferty, 2007).

Abstracts were lowercased, stripped of punctuation, numbers and publisher boilerplate, cleared of English stopwords, and stemmed; terms appearing in fewer than five documents were removed. The number of topics was selected by estimating models for K = 3 to 12 and comparing held-out likelihood, residual dispersion, the variational lower bound, and semantic coherence, which measures how often the leading terms of a topic co-occur in the same documents (Mimno et al., 2011). For the M highest-probability terms of topic k, with D(v) document frequency of term v and D(v, v′) the number of documents containing both,

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} \log[(D(v_i, v_j) + 1) / D(v_j)] \quad (3)$$

We then chose the value of K at which coherence remains high while the likelihood curves flatten. Estimation used spectral initialisation, which is deterministic given the data (Arora et al., 2013). Topics were interpreted from highest-probability terms, terms selected by frequency and exclusivity (FREX), which weight the two against each other (Bischof and Airoldi, 2012),

and the two documents most associated with each topic, with labels assigned independently by two authors and reconciled.

Prevalence was specified as a smooth function of publication year and time trends estimated by the method of composition (Roberts et al., 2019), which propagates uncertainty in the topic proportions into the standard errors. Citation impact by topic was estimated rather than tabulated as group means, since comparing means across hard-assigned topics discards partial membership and confounds topic with document age; for document i with citation count C_i, age A_i and topic proportions θ_{ik}, a negative binomial regression

$$C_i \sim \mathrm{NB}(\mu_i, \varphi), \quad \log \mu_i = \alpha + \delta \log A_i + \Sigma_{k \neq r} \lambda_k \theta_{ik} \qquad (4)$$

with topic r as reference was fitted (Cameron and Trivedi, 2013). Stability was assessed by re-estimating from 20 random initialisations and by matching topics to their closest counterparts at $K - 1$ and $K + 1$ by cosine similarity of the topic-word distributions.

2.5 Content coding and equity measures

Three dictionaries were applied to the title, abstract and keywords of each document. The country dictionary covers all 54 African states, with regular expressions written to prevent the confusion a plain string match produces, so that Sudan does not match South Sudan, Guinea does not match Equatorial Guinea or Guinea-Bissau, Congo does not match the Democratic Republic of the Congo, and Niger does not match Nigeria or the Niger Delta. Affiliations were deliberately excluded from this field, so the coded country is the country studied rather than the country an author works in. The disease dictionary covers 24 categories, and the technology dictionary includes 20, reported for the corpus, by period and by topic, which gives the entry point of each technology into the literature.

The equity measures follow from the same coding. For each African country, we report the number of documents studying it and the number of author affiliations recorded for it, and define the local authorship rate as the proportion of documents studying a country that include at least one author affiliated to an institution there. This extends the affiliation-based approach of the authorship-parasitism literature by coding the country studied independently of the country of affiliation, so that the two can be compared directly. Dictionary coding is reproducible and applies one rule to every document. Its recall nevertheless depends on the terms included, and a mention is not weighted by how central the topic is to the study.

2.6 Software and reproducibility

Analyses were conducted in R (R Core Team, 2025) using bibliometrix (Aria et al., 2017), quanteda (Benoit et al., 2018), stm (Roberts et al., 2019), strucchange (Zeileis et al., 2002) and MASS packages. A single random seed was set for the pipeline, and the search date, vocabulary threshold and seed are reported because each affects the numerical results. Every table and figure is produced by script from the screened corpus without manual intervention.

3. Results

3.1 The corpus and its growth

In Table 4, of 563 eligible documents published between 1993 and 2025 in 232 sources, 23,350 references and 12,067 citations were extracted with a median of 10 citations per document (IQR 3 to 23; mean, 21.4; SD, 40.5). Most publications were primary research: 479 articles (85.1%) and 84 reviews (14.9%). Output grew at 14.9% per year across the period. The growth was smooth rather than episodic. A supremum F test scanning every candidate break date in the log annual series returns no significant break (supF = 3.73, p = 0.798), and neither of the two dates the policy narrative would predict produces a detectable shift. The apparent acceleration visible in the raw counts, from a median of 2 documents a year before 2014 (IQR 1 to 6) to 23 in 2014 to 2019 (IQR 18 to 30) and 57 from 2020 (IQR 56 to 68), is what steady exponential growth produces: at this rate output doubles roughly every five years. Neither the West African Ebola pandemic nor the COVID-19 pandemic changed the rate at which the literature accumulates. Across the three periods before 2014, 2014 to 2019 and 2020 onwards, the number of distinct sources rose from 38 to 88 to 167, and the number of African countries appearing in the corpus rose from 20 to 33 to 46, respectively.

Table 4. Main information about the corpus

Indicator	Value	Note
Timespan	1993–2025	33 years
Documents	563	479 articles, 84 reviews
Sources	232	10 journals publish 32% of output
Annual growth rate	14.9%	no significant break (p = 0.798)
Mean document age	6.1 years	median 5, IQR 2 to 9; SD 5.0; 64.5% published from 2020
References	23,350	mean 41 per document
Citations	12,067	median 10 per document, IQR 3 to 23; mean 21.4, SD 40.5
Authors	4,299	90.6% contributed one document
Median authors per document	7	IQR 5 to 11; mean 8.6, SD 6.6; 11 single-authored documents
Degree of collaboration	0.98	collaboration index 8.78
Author keywords	1,373	present in 71% of documents
Keywords Plus	3,738	–

Table 5 gives the year-by-year counts of publications. The distribution is heavily weighted towards the present: 14 documents predate 2007 and 43 documents predate 2013, against 363 from 2020 onwards. The early records are also of a different kind. All 14 documents published before 2007 concern geospatial or remote-sensing methods rather than digitised reporting systems; electronic integrated disease surveillance and electronic records first appear in 2007, mobile tools in 2011 and district health information software in 2013 (Appendix: Figure A1). The literature therefore spans three decades, but the digitisation of surveillance systems enters it only in the second. Six years in the period produced no output at all: 1994 to 1997, 2000 and 2001, and the mean document age is 5 years (IQR 2 to 9; mean 6.1, SD 5.0).

Table 5. Annual publication output, 1993 to 2025

Year	Documents	Cumulative	% of total	Citations	Mean citations	Sources
1993	1	1	0.2	87	87.0	1
1998	1	2	0.2	281	281.0	1
1999	1	3	0.2	39	39.0	1
2002	3	6	0.5	123	41.0	3
2003	1	7	0.2	21	21.0	1
2004	1	8	0.2	32	32.0	1
2005	1	9	0.2	0	0.0	1
2006	5	14	0.9	135	27.0	5
2007	6	20	1.1	385	64.2	6
2008	2	22	0.4	85	42.5	2
2009	3	25	0.5	107	35.7	3
2010	7	32	1.2	250	35.7	6
2011	10	42	1.8	601	60.1	10
2012	1	43	0.2	13	13.0	1
2013	9	52	1.6	267	29.7	9
2014	26	78	4.6	1,492	57.4	23
2015	19	97	3.4	807	42.5	13
2016	15	112	2.7	514	34.3	13
2017	31	143	5.5	1,089	35.1	27
2018	17	160	3.0	344	20.2	15
2019	40	200	7.1	1,161	29.0	33
2020	37	237	6.6	909	24.6	30
2021	55	292	9.8	1,150	20.9	44
2022	57	349	10.1	852	14.9	43
2023	57	406	10.1	574	10.1	37
2024	71	477	12.6	491	6.9	50
2025	86	563	15.3	258	3.0	63
Total	563	–	100.0	12,067	21.4	232

Six years produced no output: 1994 to 1997, 2000 and 2001. The mean citation column falls steeply after 2020 because recent documents have had less time to accumulate citations, not because their influence is lower; 1998 shows a mean of 281 because that year contains a single, heavily cited document. The final column shows the number of distinct sources publishing in each year, which rises from one to 63 and is the clearest single indicator of the field broadening.

3.2 Sources and Leading journals

From Figure 3, sources are concentrated at the top and dispersed below it. Ten journals, 4.3% of the total, publish 32% of the output, while 169 journals, 72.8% of the total, publish one-third between them.

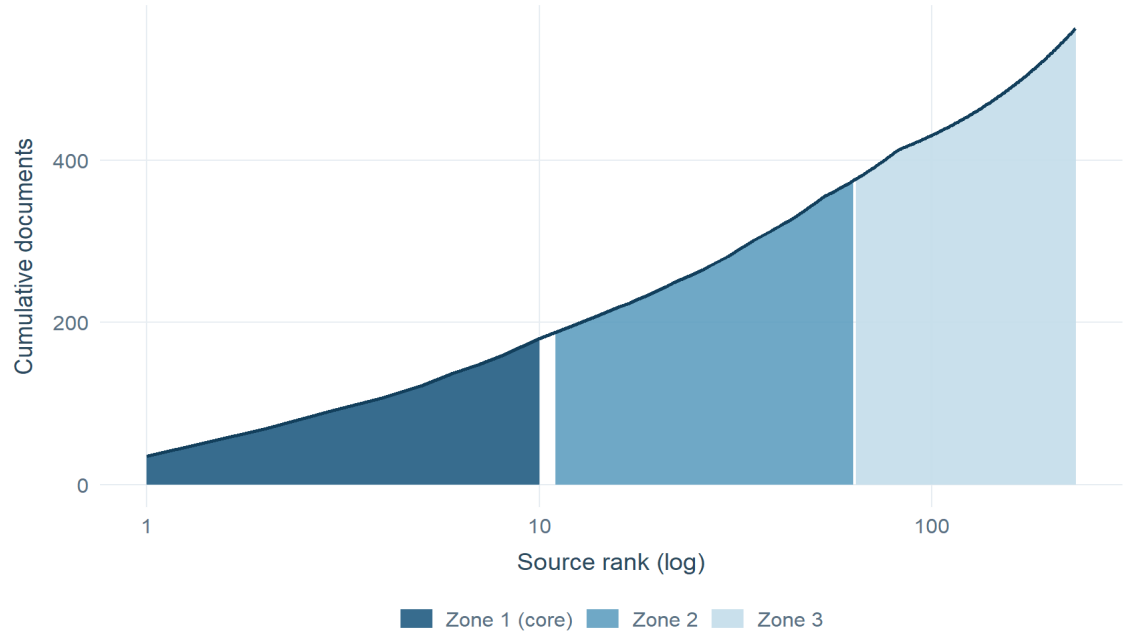


Figure 3. Bradford's law of source concentration, cumulative output by source rank.

In Table 6, Malaria Journal and PLOS ONE lead on both volume and impact, with h-indices of 16 and 14 within the corpus. No single journal accounts for more than 6% of the corpus.

Table 6. Leading sources by h-index

Source	h	g	Citations	Documents
Malaria Journal	16	25	680	34
PLOS ONE	14	22	513	35
BMC Public Health	10	15	714	15
Journal of Infectious Diseases	10	11	254	11
PLOS Neglected Tropical Diseases	9	15	248	15
BMC Infectious Diseases	8	15	382	15
Frontiers in Veterinary Science	7	7	124	7
Pan African Medical Journal	7	10	129	23

From Table 7, the United States is the largest producer by corresponding author with 115 documents, more than twice South Africa's 43, followed by the United Kingdom (34), Nigeria (32), Uganda (23), Tanzania (22), Ghana (20) and Kenya (18). Citation returns are markedly uneven: documents with German (57.7), Swiss (58.3) and British (38.3) corresponding authors average three to seven times the citations of those with Nigerian (8.2) or Ghanaian (7.4) corresponding authors. Institutionally, national ministries of health lead with 50 documents, followed by the United States Centres for Disease Control and Prevention (36) and the World Health Organisation (35); the University of Ibadan (28) is the most productive African university, ahead of Cape Town (23), Makerere (20) and Witwatersrand (18).

Table 7. Leading corresponding-author countries, with single- and multiple-country publications

Country	Articles	Citations	SCP	MCP	MCP ratio	Citations per article
United States	115	2,437	37	78	0.678	21.2
South Africa	43	732	20	23	0.535	17.0
United Kingdom	34	1,302	6	28	0.824	38.3
Nigeria	32	264	15	17	0.531	8.2
Uganda	23	215	13	10	0.435	9.3
Tanzania	22	283	8	14	0.636	12.9
Ghana	20	148	6	14	0.700	7.4
Kenya	18	477	3	15	0.833	26.5
France	15	155	3	12	0.800	10.3
Germany	15	865	1	14	0.933	57.7
Ethiopia	14	296	11	3	0.214	21.1
Switzerland	9	525	2	7	0.778	58.3
Congo	8	87	1	7	0.875	10.9
Mozambique	8	28	5	3	0.375	3.5

SCP, single-country publications; MCP, multiple-country publications. Ranked by articles; the twenty most productive countries are in the analysis output.

3.3 Intellectual base and research front

In the Appendix, Figure A7, the literature was markedly non-cumulative. Only 149 of 563 documents (26.5%) were cited even once by another document in the corpus, and the most locally cited paper receives nine such citations. Fifty-nine documents reach the threshold of two local citations required to construct a historical citation network, so the field's internal citation structure is thin enough that its lineage cannot be traced through the usual method.

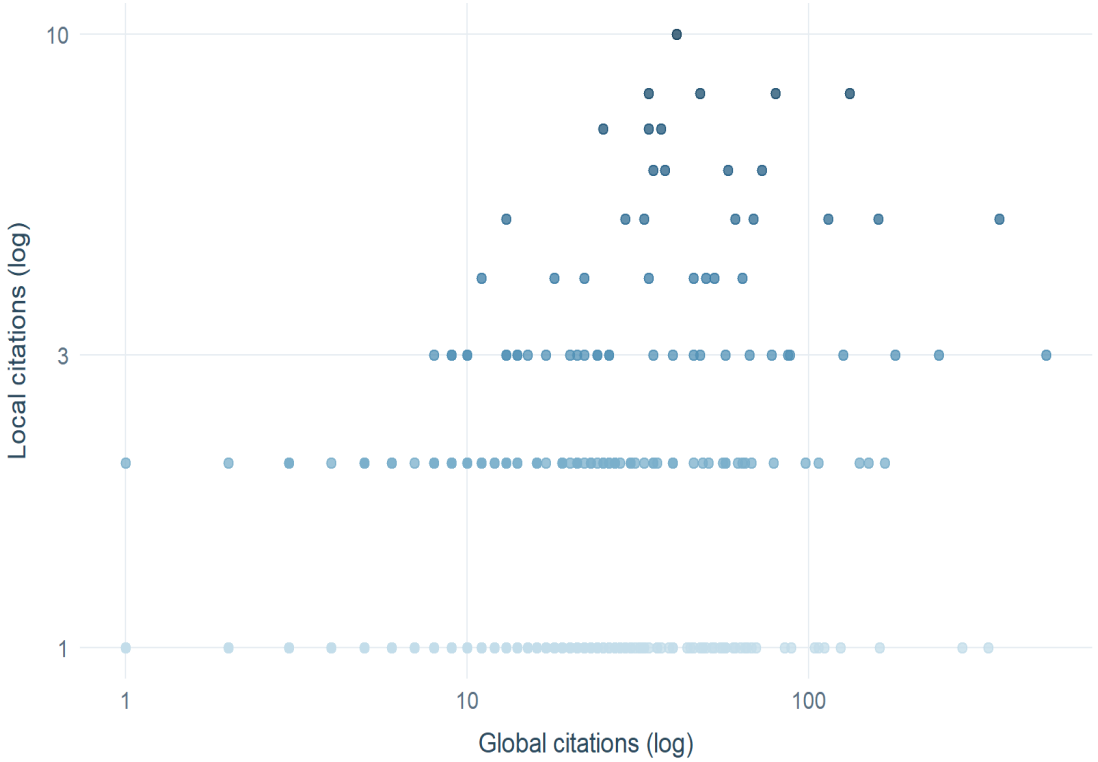


Figure 4. Local against global citation scores. Points low and to the right are cited widely outside the corpus but rarely within it; the mass of documents on the lower axis receive no local citations at all.

Table 8 presents the most co-cited references. These co-cited references are not research articles but institutional documents: WHO guidance on the COVID-19 pandemic, and the World Malaria Report (WHO, 2005). The Price index as given in Figure 5 is the share of each document's references published within the preceding five years is 0.531 overall and there is stability over decades (0.486, 0.504, 0.536, 0.530 from the 1990s to the 2020s), indicating a literature that draws consistently on recent works rather than on a canon. Cited-reference years peak sharply in 2020 at more than six times the next largest deviation, reflecting the volume of COVID-19 pandemic works published and cited within a single year rather than a foundational contribution.

Table 8. Most co-cited references

Reference	Type
WHO, Coronavirus disease (COVID-19)	Institutional guidance
WHO, World Malaria Report 2019	Institutional report
International Health Regulations (2005)	Treaty instrument
WHO, Coronavirus disease 2019 (COVID-19)	Institutional guidance
Nkengasong JN, 2020	Research article
Potential impact of health service disruptions on the burden of malaria, 2020	Research article

Four of the six most co-cited references are institutional documents rather than research articles. This is the clearest expression of the pattern described in section 4.2 of the main text: the field's shared foundations are policy instruments, not a scholarly canon.

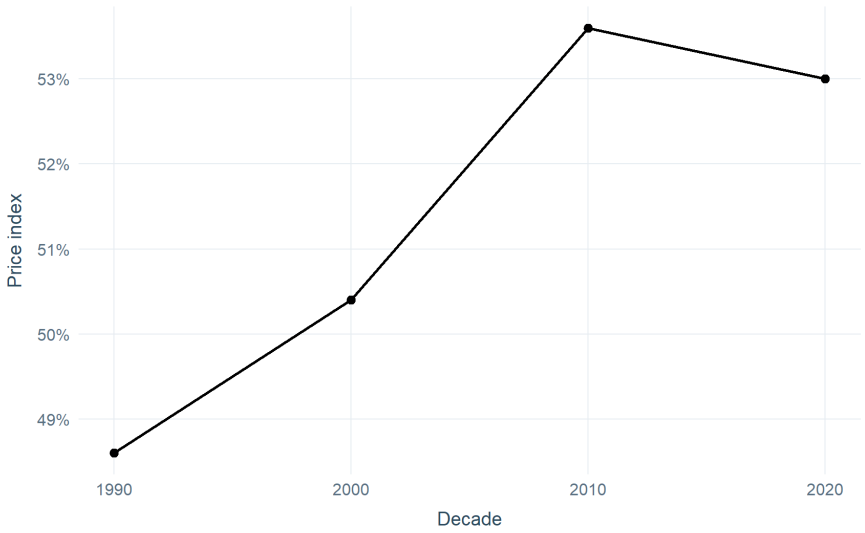


Figure 5. Price index by decade: the share of references published within five years of the citing document.

Bibliographic coupling identifies a research front organised around six clusters (Appendix: Figure A6), the largest containing 34 documents and centred on community-based One Health surveillance in Tanzania. That the front is more coherent than the intellectual base is consistent with a field in which recent work shares references without yet citing itself.

3.4 Thematic structure

Diagnostics supported a six-topic solution (Appendix: Figure A2). Table 9 reports the topics, their terms, prevalence, quality and time trend. Semantic coherence ranges from −29.7 to −65.5 and exclusivity from 8.69 to 9.60, with no topic collapsing towards either. A co-word thematic map built independently from author keywords recovers a comparable structure, with surveillance and system strengthening central and the animal and geospatial themes peripheral (Appendix: Figure A3). The model was stable across 20 random initialisations (coherence −50.5 to −55.1, exclusivity 9.20 to 9.29; Appendix: Figure A4), and five of the six topics have a counterpart at K = 5 with cosine similarity above 0.95 (Appendix: Table A1).

Table 9. The six topics, with prevalence, quality, citation impact and time trend

No.	Label	Distinctive terms (FREX)	Prev.	Docs	Coh.	Excl.	Slope	p
5	IDSR system strengthening and response capacity	idsr, respons, strengthen, capac	0.224	135	−29.7	8.74	+0.0017	0.490
6	Digital health innovation and data sharing	technolog, innov, share, mhealth	0.179	101	−50.1	8.69	+0.0087	<0.001
2	Animal and zoonotic pathogen surveillance	anim, asf, swine, pathogen	0.172	96	−65.5	9.22	−0.0008	0.750
3	Clinical and programme data for HIV care	hiv, patient, care, clinic	0.152	96	−60.1	9.16	−0.0054	0.023
4	Routine malaria reporting at district level	malaria, incid, case, district	0.140	64	−58.5	9.42	−0.0025	0.220
1	Geospatial risk mapping and vector surveillance	spatial, cluster, map, mosquito	0.134	71	−61.3	9.60	−0.0017	0.410

Prev., expected topic prevalence; Docs, documents assigned by highest probability; Coh., semantic coherence; Excl., exclusivity. Terms are stemmed, as produced by the model. Slopes are from estimateEffect on publication year with global uncertainty; the document-level and n-weighted annual estimators agree to the third decimal.

Two topics moved significantly. Digital health innovation and data sharing showed a significant rise (+0.0087 per year, $p < 0.001$), while clinical and programme data for HIV care demonstrated significant decline (−0.0054, $p = 0.023$). The other four topics were flat. Notably, the IDSR system strengthening was the most prevalent topic (0.224) and the most internally coherent (−29.7), yet it showed no trend. This suggests that the system-strengthening literature has grown in step with the field rather than faster or slower than it.

Themes also separate by venue: the leading journals carry markedly different topic mixtures, with Malaria Journal dominated by routine malaria reporting and Frontiers in Veterinary Science by animal surveillance (Figure 6). Topics were strongly and negatively correlated, most sharply between geospatial mapping and IDSR system strengthening (r = −0.415), between animal surveillance and clinical data (−0.285) and between animal surveillance and malaria reporting (−0.274). Documents in this field therefore tend to belong to one theme rather than bridging several (Table 10), which is a measure of how separated its research communities are.

Table 10. Correlations between topic proportions

Topic	T1	T2	T3	T4	T5	T6
T1 Geospatial mapping	1.000	0.028	−0.174	−0.070	−0.415	−0.200
T2 Animal and zoonotic	0.028	1.000	−0.285	−0.274	−0.227	−0.182
T3 Clinical HIV data	−0.174	−0.285	1.000	−0.102	−0.282	−0.185
T4 Malaria reporting	−0.070	−0.274	−0.102	1.000	−0.158	−0.349
T5 IDSR strengthening	−0.415	−0.227	−0.282	−0.158	1.000	−0.101
T6 Digital innovation	−0.200	−0.182	−0.185	−0.349	−0.101	1.000

Estimated from the unrestricted covariance of the logistic-normal prior. No pair is substantially positively correlated.

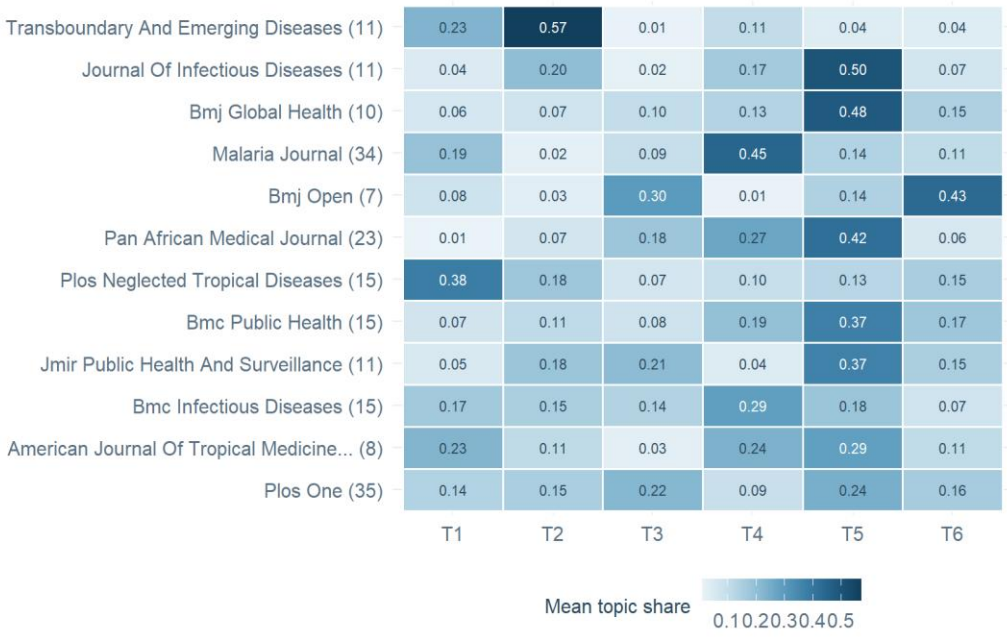


Figure 6. Mean topic proportion across the leading journals.

In Table 11, the citation impact differs by topic after adjusting for the number of years each document has been available to be cited. In a negative binomial model with the logarithm of document age as a covariate, and geospatial mapping as the reference, documents in the digital innovation topic attract 2.05 times the citations (p = 0.007), while clinical data (0.48, p = 0.005), malaria reporting (0.54, p = 0.027) and IDSR system strengthening (0.45, p < 0.001) attracted

roughly half. Age dominates all of it: each unit increase in log age multiplies expected citations by 3.02. The topic with the most documents is thus the least cited per document, and the fastest-growing topic is the most cited.

Table 11. Negative binomial model of citation counts by topic, with document age controlled

Term	Coefficient	Multiplier	p
Intercept	1.351	3.861	<0.001
log(document age)	1.107	3.024	<0.001
T2 Animal and zoonotic surveillance	−0.245	0.783	0.395
T3 Clinical and programme data for HIV	−0.728	0.483	0.005
T4 Routine malaria reporting	−0.623	0.536	0.027
T5 IDSR system strengthening	−0.805	0.447	<0.001
T6 Digital health innovation	0.715	2.045	0.007

Geospatial risk mapping is the reference topic. The multiplier is the change in expected citations for a document wholly in that topic rather than wholly in the reference, holding age constant.

3.5 Coverage of countries, diseases and technologies

Figure 7 shows that six African countries are of particular interest. Nigeria and South Africa are each studied in 61 documents, Uganda in 58, Kenya in 50, Tanzania in 48, and Ethiopia and Ghana in 29 each; together these account for the majority of country-specific work. The distribution then falls away sharply: 15 documents concern the Democratic Republic of the Congo, 13 Burkina Faso and Guinea, and eight or fewer concern Niger, Chad, Somalia and most of the remainder.

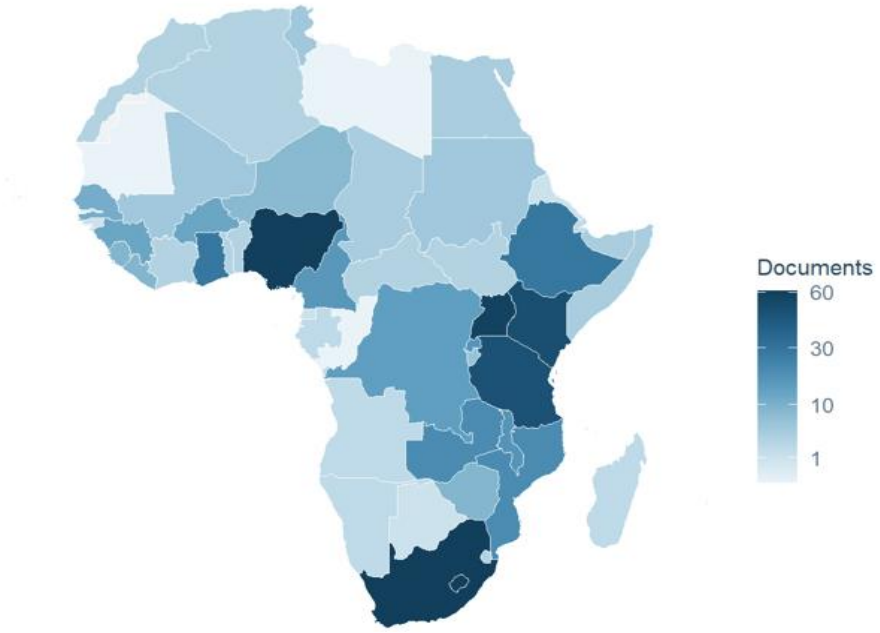


Figure 7. Documents by African country studied, coded from title, abstract and keywords rather than author affiliation.

In Table 12, disease coverage tracks the continent's established priorities more than its emerging ones. Malaria appears in 107 documents (19.0%), COVID-19 pandemic in 87

(15.5%), HIV in 55 (9.8%), tuberculosis and Ebola in 44 each (7.8%), and influenza in 39 (6.9%). Antimicrobial resistance, the neglected tropical diseases and non-communicable disease each appear in under 4% of documents, and mpox in 15 (2.7%) despite the 2024 continental emergency. Attention to COVID-19 pandemic rose from four documents before 2020 to 83 after it, while attention to malaria, HIV and tuberculosis grew in proportion to the field.

Table 12. The twelve most frequent diseases and surveillance technologies

Disease	Documents	%	Technology	Documents	%
Malaria	107	19.0	Geospatial analysis and GIS	109	19.4
COVID-19	87	15.5	Machine learning and artificial intelligence	84	14.9
HIV and AIDS	55	9.8	Mobile health (mHealth)	62	11.0
Tuberculosis	44	7.8	DHIS2	47	8.3
Ebola	44	7.8	Integrated disease surveillance	40	7.1
Influenza	39	6.9	Electronic health records	39	6.9
Maternal and child health	27	4.8	Smartphone applications	38	6.7
Cholera	22	3.9	Telemedicine	31	5.5
Meningitis	22	3.9	Remote sensing	30	5.3
Measles	18	3.2	Social media and internet	27	4.8
Polio	18	3.2	Genomic surveillance	25	4.4
Non-communicable disease	18	3.2	Interoperability and standards	15	2.7

Percentages are of the 563 documents; a document may address more than one disease or technology.

Technology coverage was dominated by methods rather than systems. Geographic information systems (GIS) and other geospatial analysis appear in 109 documents (19.4%) and machine learning and artificial intelligence in 84 (14.9%), ahead of mobile health tools (62, 11.0%), DHIS2 (47, 8.3%) and integrated disease surveillance platforms (40, 7.1%). Interoperability and data standards, the requirement most implementation reviews identify as the binding constraint, appear in 15 documents (2.7%), even though “information system” and its variants are among the most frequent terms in the corpus at every n-gram length (Appendix: Table A2); wastewater surveillance in four and unmanned aerial systems in three. Short message service (SMS) reporting illustrates a technology that entered and did not scale: five documents in 2014 to 2019 and five in the six years since, against 26 for smartphone applications.

The entry dates in Figure A1 (in the Appendix) indicate that the field adopts each new method roughly as it becomes available, with two exceptions. Genomic surveillance entered in 2015 but reached a tenth of annual output only in 2024, and machine learning entered in 2016 but not until 2021. Dashboards, interoperability, event-based surveillance, participatory reporting, SMS reporting and wastewater monitoring have never reached that threshold in any year.

3.6 Attention and local authorship

From Figure 8, across the 35 African countries appearing in at least three documents, the proportion of studies of a country that include an author affiliated there rises with the attention that country receives (Spearman $\rho = 0.598$, $p < 0.001$; ordinary least squares on log documents, slope 0.128, $p = 0.001$, $R^2 = 0.27$). Attention and local author presence are still more tightly coupled ($\rho = 0.839$, $p < 0.001$).

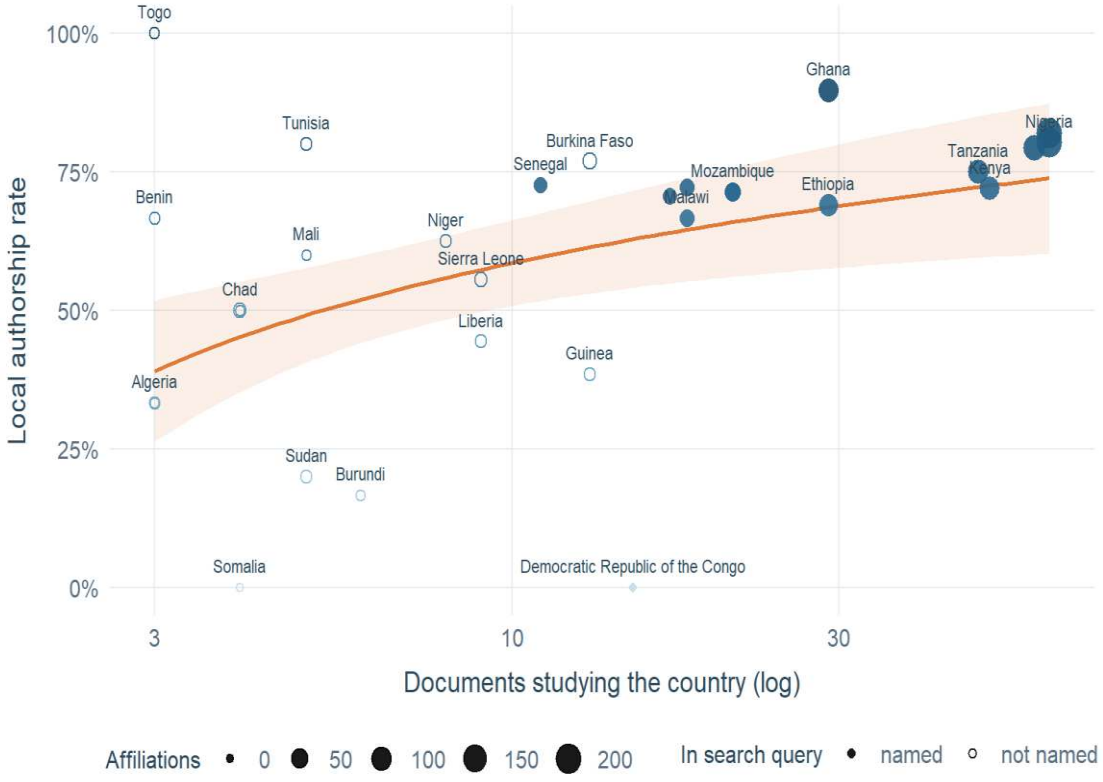


Figure 8. Documents studying each country against the local authorship rate, with a fitted logarithmic trend. Hollow points mark countries not named individually in the search query.

In Table 13, it is apparent that the inequity is not uniform across the continent but concentrated in the countries that receive the least attention. Ghana (0.897), South Africa (0.820), Nigeria

(0.803), Uganda (0.793), Tanzania (0.750) and Kenya (0.720) combine close attention with high local participation. At the other end, the Democratic Republic of the Congo is studied in 15 documents with no author affiliation recorded in the country at all, a local authorship rate of zero; Guinea stands at 0.385, Liberia at 0.444, and Sierra Leone and Zimbabwe at 0.556. Based on Table 14, countries that were named individually in the search query have significantly higher local authorship than those that were not (Wilcoxon W = 241, p = 0.002), which is consistent with this pattern but also, as Section 2.2 notes, partly a property of the search.

Table 13. Attention and local authorship, selected countries

Country	Documents studying	Author affiliations	Local authorship rate
Ghana	29	96	0.897
South Africa	61	235	0.820
Nigeria	61	230	0.803
Uganda	58	137	0.793
Tanzania	48	115	0.750
Kenya	50	113	0.720
Ethiopia	29	81	0.690
Niger	8	9	0.625
Sierra Leone	9	15	0.556
Liberia	9	8	0.444
Guinea	13	11	0.385
Democratic Republic of the Congo	15	0	0.000

Table 14. Tests of the relationship between attention and local authorship

Test	Statistic	p	n
Spearman: documents studying vs local authorship rate	ρ = 0.598	<0.001	35
Spearman: documents studying vs author affiliations	ρ = 0.839	<0.001	35
Wilcoxon: local authorship rate, query-named vs not named	W = 241	0.002	35
OLS: local authorship rate on log(documents studied)	slope 0.128	0.001	35
OLS: R-squared	0.274	–	35

Restricted to the 35 African countries appearing in at least three documents.

4. Discussion

Three findings run against what the policy narrative would predict, and together they describe a field whose difficulties are structural rather than a matter of volume.

4.1 Growth without inflexion

Research on digital disease surveillance in Africa has grown at a near-constant 15% a year across the whole period, through the West African Ebola and COVID-19 pandemics, and the mpox emergency, without a statistically detectable break at any of them. Pandemics are routinely described as inflexion points for surveillance research (Tom-Aba et al., 2018; Tegally et al., 2022; Africa Centres for Disease Control and Prevention, 2024), and the raw annual counts appear to support that reading. They do not survive a formal test. What pandemics appear to have changed is the composition of the field rather than its rate: the number of sources publishing in it quadrupled, and the number of countries appearing in it more than doubled, while the growth rate held. The most plausible reading is that output is governed by research capacity, which accumulates steadily, rather than by attention, which spikes.

4.2 A field that does not cite itself

Only a quarter of these documents are cited even once by another document in the corpus, and the most locally cited paper attracts nine such citations. The internal citation network is too sparse to support a historical lineage analysis. Meanwhile, the most co-cited references are not peer-reviewed research outputs at all, but WHO guidance, the World Malaria Report and the International Health Regulations, and the Price index has sat close to 0.53 in every decade for which there are enough documents to measure it, indicating a literature that reads recent work rather than a canon.

This is consistent with the implementation picture established elsewhere. A systematic review of digitalisation of infectious disease surveillance in Africa found 29 reports describing 27 projects in 13 countries, three-quarters of them pilots with a median life of sixteen months (Kaburi et al., 2024). A field of short, unconnected pilots would produce exactly this citation structure: each project reports its own experience, cites the policy framework it operates under, and does not build on the project before it. The negative topic correlations point the same way, with no pair of themes substantially positively correlated, indicating research communities that share a subject but not a conversation.

The practical implication is that the volume of this literature overstates its cumulative value. Five hundred and sixty-three documents growing at 15% a year is a healthy-looking field; a field in which three quarters of the work is never cited by its own successors is not accumulating knowledge at anything like that rate.

4.3 Coverage misaligned with need

Two mismatches stand out. The first is thematic. Geospatial analysis and machine learning together appear in a third of the corpus, while interoperability and data standards appear in 2.7%. Implementation reviews consistently identify fragmentation and the absence of interoperability as the binding constraint on digital surveillance in the region (Kaburi et al., 2024; Mremi et al., 2021), and a scoping review of digital technologies for infectious disease surveillance in Tanzania found most deployed technologies to be stand-alone (Mustafa et al., 2023). The research literature is concentrated on analytical methods rather than on the integration that would allow systems to exchange information. Dashboards, event-based surveillance, participatory reporting and interoperability have never reached a tenth of annual output in any year of the period.

The second is epidemiological. Mpox appears in 15 documents despite the first continental public health emergency declared over it, and antimicrobial resistance, neglected tropical diseases and non-communicable disease each in under 4%. The disease profile of this literature tracks the established vertical programmes for malaria, HIV and tuberculosis, and the newest continental emergency is close to absent from it.

4.4 Equity is a question about neglected countries

Our clearest departure from the wider authorship literature is that local participation and attention move together rather than against each other. Countries studied more often are more likely to have local authors in the studies about them, not less. The six best-covered countries all have local authorship rates between 0.72 and 0.90, which is higher than the figures reported for African health research generally (Hedt-Gauthier et al., 2019; Mbaye et al., 2019; Kisenge et al., 2021).

That does not mean the equity concern is unfounded. It means it is concentrated. Where a country's surveillance research is thin, it is also disproportionately conducted from outside: the Democratic Republic of the Congo, a country of over 100 million people with recurrent Ebola and cholera outbreaks, is studied in 15 documents and appears as an author affiliation in none of them. Guinea, Liberia and Sierra Leone, the three countries of the 2014 to 2016 epidemic

(Tom-Aba et al., 2018), sit between 0.39 and 0.56. The pattern is not extraction from a continent but a gap that widens as capacity thins, so that the countries least able to generate their own surveillance evidence are also those most likely to be written about by others. That is a more precise, and more actionable, statement than the general finding, because it identifies where research capacity investment would change the picture.

4.5 Strengths and limitations

The analysis is scripted end to end from a screened corpus with stated eligibility criteria, a recorded search date and a fixed seed, and the code and corpus are available, so every figure can be reproduced. Two tests were pre-specified rather than selected after inspection: the break tests at 2014 and 2020, and the tests of the attention-authorship relationship.

Four limitations bear on interpretation. First, the search named only fourteen African countries individually alongside the continental terms, so documents about the other forty entered the corpus only if they also used a continental term. Among documents about named countries, 54% use no continental term, implying that counts for unnamed states understate their literature by a factor of roughly 2.2. The country coverage results should be read as the distribution of attention within this corpus, not within the world literature, and the difference between named and unnamed countries in local authorship is partly attributable to this. Second, the single-database, English-only design omits African Journals Online and the French and Portuguese literature; 173 retained documents concern francophone countries but none is francophone in language. Third, residual dispersion in the topic model is 1.60 with $p < 0.001$, indicating that more than six topics could be justified; the six re ported here are stable and interpretable but coarse. Fourth, coded country attention counts mentions rather than weighting them by centrality to the study, and a document naming five countries contributes to five counts.

5. Conclusion

Research on digital disease surveillance in Africa has grown steadily since 1993 without inflecting at any of the pandemics that were supposed to transform it, and it has grown broader rather than deeper. Three-quarters of it is never cited by its own successors, its most co-cited sources are policy documents rather than research, and it concentrates on analytical methods rather than on the interoperability that implementation reviews identify as the constraint. Attention is held by six countries, and where attention is thinnest, local authorship is thinnest too, most starkly in the Democratic Republic of the Congo, studied in fifteen documents and authored in none.

Two implications follow for those who fund and publish this work. Continuity matters more than volume: a field of short pilots that do not cite each other will not accumulate the evidence base that surveillance policy needs, however fast it grows. And research capacity investment is best targeted where the attention-authorship gap is widest, which is not the continent as a whole but an identifiable set of fragile and conflict-affected states that are studied from outside when they are studied at all.

Declarations

Author contributions. CA conceived the study, assembled and screened the corpus, and performed the bibliometric and structural topic modelling analyses. CA and OSY drafted the manuscript. RFA and DOO contributed to editing and revising the manuscript. All authors read and approved the final version.

Funding. The authors received no specific funding for this work.

Competing interests. The authors declare no competing interests.

Ethical approval. Not applicable.

Consent to participate. Not applicable.

Consent to publish. Not applicable.

Data availability. Not applicable.

References

Adokiya MN, Awoonor-Williams JK, Beiersmann C, Muller O. The integrated disease surveillance and response system in northern Ghana: challenges to the core and support functions. BMC Health Services Research. 2015;15:288.

Africa Centres for Disease Control and Prevention. Africa CDC declares mpox a public health emergency of continental security [press release]. Addis Ababa: Africa CDC; 13 August 2024.

African Union. The African Union launches Africa CDC, a continent-wide public health agency [press release]. Addis Ababa: African Union Commission; 31 January 2017.

Agbeyangi AO, Agrawal S, Lukose JM. Digital health implementation research across selected African countries: a bibliometric analysis of maternal health and infectious diseases with observations on precision medicine representation (2015-2025). Frontiers in Digital Health. 2026;8:1831353.

Aku F, Amuasi JH, Debrah LB, Opoku D, Gmanyami JM, Debrah AY, et al. mHealth tools for community-based infectious disease surveillance in Africa: a scoping review. European Journal of Public Health. 2023;33(Suppl 2):ckad160.020.

Aria M, Cuccurullo C. bibliometrix: an R-tool for comprehensive science mapping analysis. Journal of Informetrics. 2017;11(4):959-75.

Arora S, Ge R, Halpern Y, Mimno D, Moitra A, Sontag D, et al. A practical algorithm for topic modeling with provable guarantees. In: Proceedings of the 30th International Conference on Machine Learning. 2013. p. 280-8.

Ayantse C, Yaya OS, Sakpere AB, Abel OKD, Ojo IO, Joseph OU. Infectious diseases in Nigeria using topic modelling: a systematic review. Journal of Advances in Mathematics and Computer Science. 2024;39(11):59-75.

Baas J, Schotten M, Plume A, Cote G, Karimi R. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies. 2020;1(1):377-86.

Bai J, Perron P. Computation and analysis of multiple structural change models. Journal of Applied Econometrics. 2003;18(1):1-22.

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Muller S, et al. quanteda: an R package for the quantitative analysis of textual data. Journal of Open Source Software. 2018;3(30):774.

Bischof J, Airoldi EM. Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th International Conference on Machine Learning. 2012. p. 201-8.

Blei DM, Lafferty JD. A correlated topic model of Science. Annals of Applied Statistics. 2007;1(1):17-35.

Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003;3:993-1022.

Bradford SC. Sources of information on specific subjects. Engineering. 1934;137:85-6.

Callon M, Courtial JP, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research. Scientometrics. 1991;22(1):155-205.

Cameron AC, Trivedi PK. Regression analysis of count data. 2nd ed. Cambridge: Cambridge University Press; 2013.

Cobo MJ, Lopez-Herrera AG, Herrera-Viedma E, Herrera F. An approach for detecting, quantifying, and visualizing the evolution of a research field. Journal of Informetrics. 2011;5(1):146-66.

Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. Journal of Business Research. 2021;133:285-96.

Hedt-Gauthier BL, Jeufack HM, Neufeld NH, Alem A, Sauer S, Odhiambo J, et al. Stuck in the middle: a systematic review of authorship in collaborative health research in Africa, 2014-2016. BMJ Global Health. 2019;4(5):e001853.

Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences. 2005;102(46):16569-72.

Ibrahim LM, Stephen M, Okudo I, Kitgakka SM, Mamadu IN, Njai IF, et al. A rapid assessment of the implementation of integrated disease surveillance and response system in Northeast Nigeria. BMC Public Health. 2020;20(1):600.

Joseph Wu TS, Kagoli M, Kaasboll JJ, Bjune GA. Integrated Disease Surveillance and Response (IDSR) in Malawi: implementation gaps and challenges for timely alert. PLOS ONE. 2018;13(11):e0200858.

Kaburi BB, Harries M, Hauri AM, Kenu E, Wyss K, Silenou BC, et al. Availability of published evidence on coverage, cost components, and funding support for digitalisation of infectious disease surveillance in Africa, 2003-2022: a systematic review. BMC Public Health. 2024;24(1):1731.

Kessler MM. Bibliographic coupling between scientific papers. American Documentation. 1963;14(1):10-25.

Kisenge R, Duggan CP, Ali M, Rees CA, Ideh RC, Sirna SJ, et al. Where there is no local author: a network bibliometric analysis of authorship parasitism among research conducted in sub-Saharan Africa. BMJ Global Health. 2021;6(10):e006982.

Lotka AJ. The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences. 1926;16(12):317-23.

Mandyata CB, Olowski LK, Mutale W. Challenges of implementing the integrated disease surveillance and response strategy in Zambia: a health worker perspective. BMC Public Health. 2017;17(1):746.
Marx W, Bornmann L, Barth A, Leydesdorff L. Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). Journal of the Association for Information Science and Technology. 2014;65(4):751-64.

Mbaye R, Gebeyehu R, Hossmann S, Mbarga N, Bih-Neh E, Eteki L, et al. Who is telling the story? A systematic review of authorship for infectious disease research conducted in Africa, 1980-2016. BMJ Global Health. 2019;4(5):e001855.

Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011. p. 262-72.

Mongeon P, Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics. 2016;106(1):213-28.

Mremi IR, George J, Rumisha SF, Sindato C, Kimera SI, Mboera LEG. Twenty years of integrated disease surveillance and response in sub-Saharan Africa: challenges and opportunities for effective management of infectious disease epidemics. One Health Outlook. 2021;3(1):22.

Mustafa UK, Kreppel KS, Brinkel J, Sauli E. Digital technologies to enhance infectious disease surveillance in Tanzania: a scoping review. Healthcare. 2023;11(4):470.

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

Perry HN, McDonnell SM, Alemu W, Nsubuga P, Chungong S, Otten MW Jr, et al. Planning an integrated disease surveillance and response system: a matrix of skills and activities. BMC Medicine. 2007;5:24.

Phoobane P, Masinde M, Mabhaudhi T. Predicting infectious diseases: a bibliometric review on Africa. International Journal of Environmental Research and Public Health. 2022;19(3):1893.

Price DJD. Citation measures of hard science, soft science, technology, and nonscience. In: Nelson CE, Pollack DK, editors. Communication among scientists and engineers. Lexington: Heath; 1970. p. 3-22.

R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2025.

Regional Committee for Africa. Integrated disease surveillance in Africa: a regional strategy (1999-2003). Report of the Regional Director. Geneva: World Health Organization; 1998.

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, et al. Structural topic models for open-ended survey responses. American Journal of Political Science. 2014;58(4):1064-82.

Roberts ME, Stewart BM, Tingley D. stm: an R package for structural topic models. Journal of Statistical Software. 2019;91(2):1-40.

Silenou BC, Nyirenda JLZ, Zaghloul A, Lange B, Doerrbecker J, Schenkel K, et al. Availability and suitability of digital health tools in Africa for pandemic control: scoping review and cluster analysis. JMIR Public Health and Surveillance. 2021;7(12):e30106.

Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the American Society for Information Science. 1973;24(4):265-9.

Tegally H, San JE, Cotten M, Moir M, Tegomoh B, Mboowa G, et al. The evolving SARS-CoV-2 epidemic in Africa: insights from rapidly expanding genomic surveillance. Science. 2022;378(6615):eabq5358.

Tom-Aba D, Nguku PM, Arinze CC, Krause G. Assessing the concepts and designs of 58 mobile apps for the management of the 2014-2015 West Africa Ebola outbreak: systematic review. JMIR Public Health and Surveillance. 2018;4(4):e68.

WHO. International Health Regulations. 3rd edt. Geneva: World Health Organization; 2005.

Wolfe CM, Hamblion EL, Dzotsi EK, Mboussou F, Eckerle I, Flahault A, et al. Systematic review of Integrated Disease Surveillance and Response (IDSR) implementation in the African region. PLOS ONE. 2021;16(2):e0245457.

Zeileis A, Leisch F, Hornik K, Kleiber C. strucchange: an R package for testing for structural change in linear regression models. Journal of Statistical Software. 2002;7(2):1-38.

Appendix

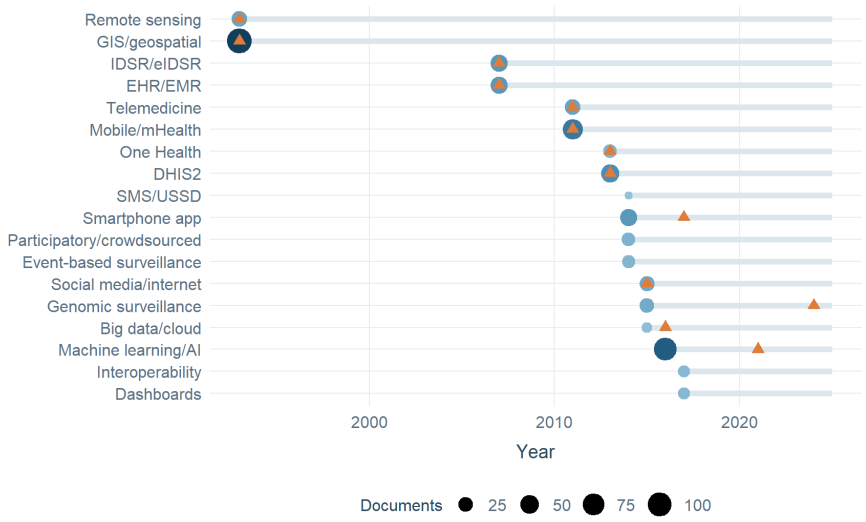


Figure A1. When each surveillance technology entered the literature. Circles mark the first year a technology appears, sized by the number of documents mentioning it; triangles mark the first year it reached ten per cent of that year's output.

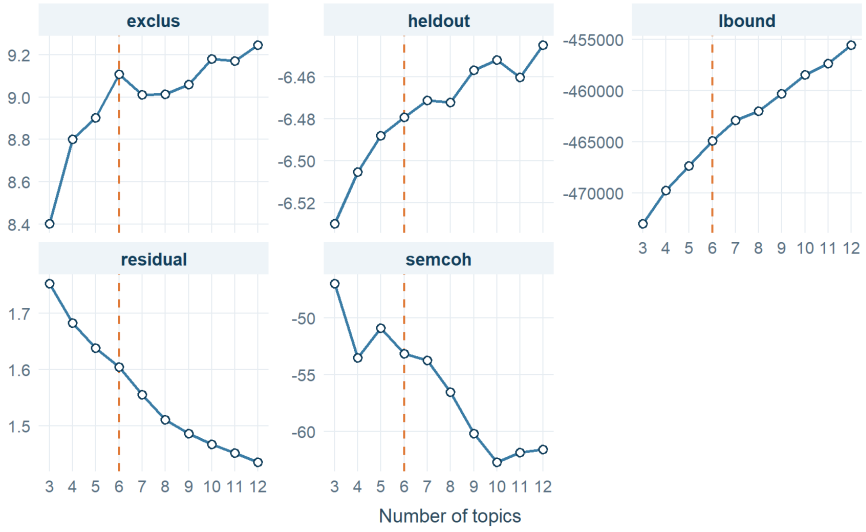


Figure A2. Diagnostics for candidate numbers of topics: held-out likelihood, residual dispersion, semantic coherence, exclusivity and the variational lower bound for K = 3 to 12.

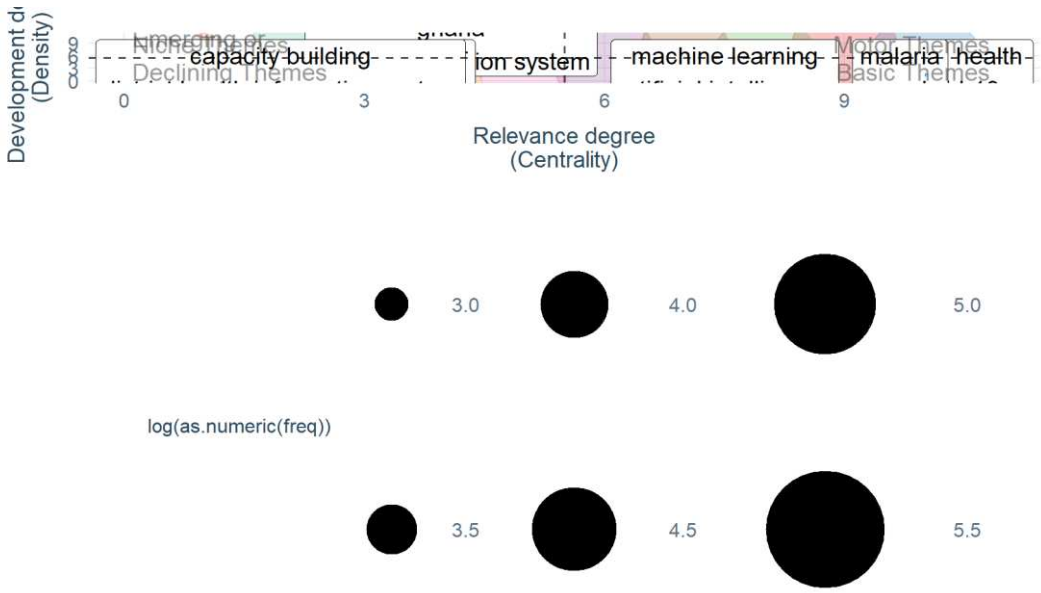


Figure A3. Thematic map of author keywords, on axes of centrality and density.

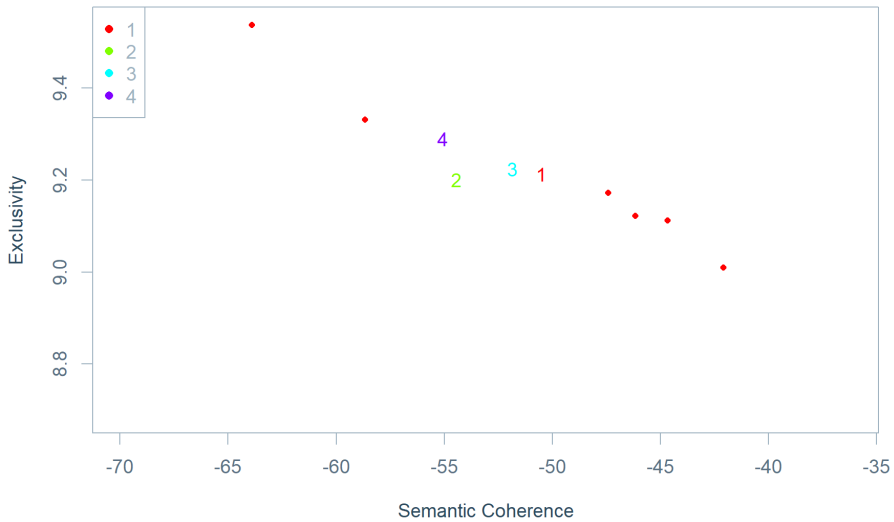


Figure A4. Semantic coherence against exclusivity across twenty random initialisations, with the spectral solution used in the main analysis marked.

Table A1. Topic stability across the number of topics

Topic at K = 6	Best match at K = 5	Cosine	Best match at K = 7 (cosine)
1 Geospatial mapping	1	0.990	1 (0.986)
2 Animal and zoonotic	2	0.993	2 (0.977)
3 Clinical HIV data	3	0.996	see note
4 Malaria reporting	4	0.953	see note
5 IDSR strengthening	5	0.891	see note
6 Digital innovation	5	0.861	see note

Cosine similarity of the topic-word distributions. Five of the six topics have a near-identical counterpart at K = 5. Topics 5 and 6 both map onto the same K = 5 topic, which indicates that the separation of system strengthening from digital innovation is the distinction the sixth topic adds.

Table A2. Most frequent terms in abstracts, by n-gram length

Unigram	n	Bigram	n	Trigram	n
health	1,593	public health	349	public health surveillance	57
data	1,460	disease surveillance	264	world health organization	42
surveillance	1,236	surveillance system	151	integrated disease surveillance	42
disease	795	surveillance systems	117	district health information	41
system	662	information system	112	health information system	38
malaria	634	health information	110	health management information	30
information	449	health facilities	105	infectious disease surveillance	28
cases	439	surveillance data	93	management information system	26
africa	418	infectious diseases	88	health information systems	25
systems	415	infectious disease	80	african swine fever	23

The prominence of “information system” and its variants across all three lengths is notable given that interoperability and data standards appear in only 2.7% of documents: the literature discusses information systems far more often than it addresses their integration.

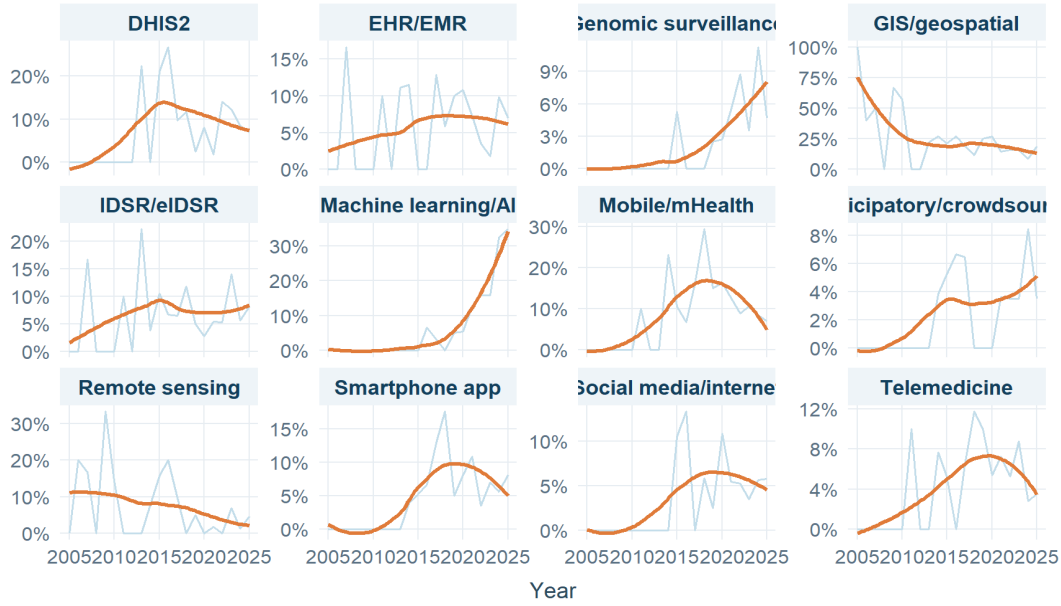


Figure A5. Technology adoption over time: share of each year's documents mentioning each technology, with a loess fit.

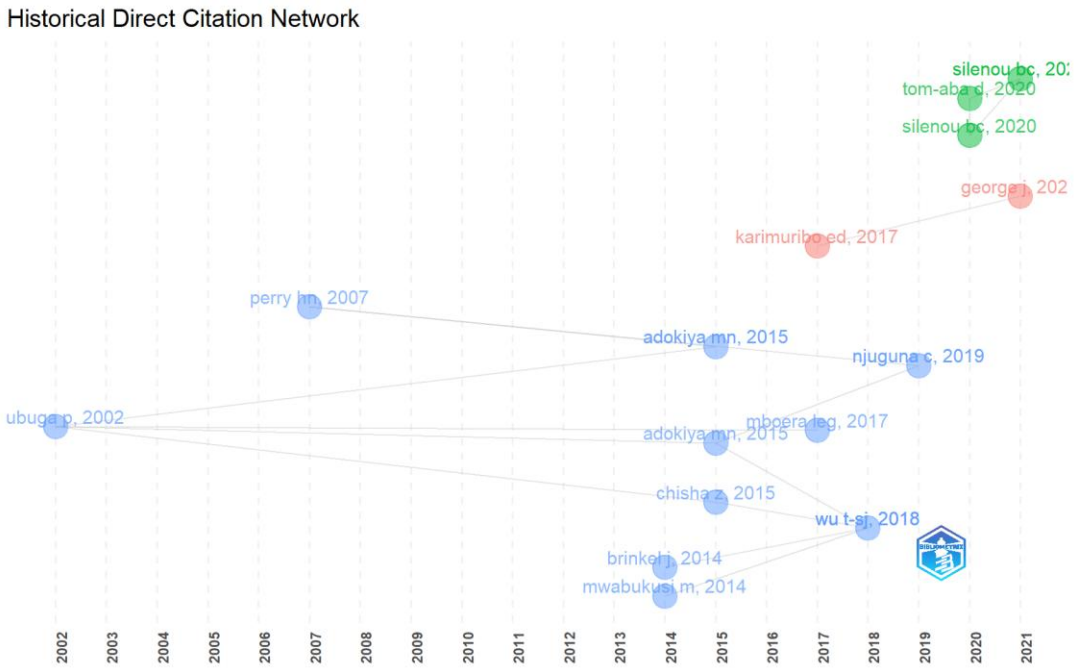


Figure A6. Bibliographic coupling map: the current research front.

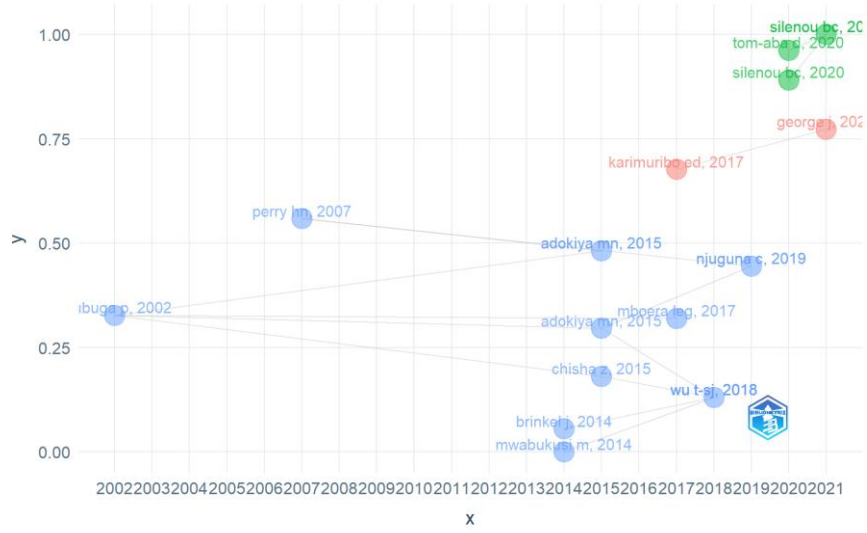


Figure A7. Historical direct citation network. Only 59 documents reach the two-local-citation threshold, so the lineage is necessarily sparse.