Secondary data for global health digitalisation

Anatol-Fiete Näher, Carina N Vorisek, Sophie A I Klopfenstein, Moritz Lehne, Sylvia Thun, Shada Alsalamah, Sameer Pujari, Dominik Heider, Wolfgang Ahrens, Iris Pigeot, Georg Marckmann, Mirjam A Jenny, Bernhard Y Renard, Max von Kleist, Lothar H Wieler, Felix Balzer, Linus Grabenhenrich

Substantial opportunities for global health intelligence and research arise from the combined and optimised use of secondary data within data ecosystems. Secondary data are information being used for purposes other than those intended when they were collected. These data can be gathered from sources on the verge of widespread use such as the internet, wearables, mobile phone apps, electronic health records, or genome sequencing. To utilise their full potential, we offer guidance by outlining available sources and approaches for the processing of secondary data. Furthermore, in addition to indicators for the regulatory and ethical evaluation of strategies for the best use of secondary data, we also propose criteria for assessing reusability. This overview supports more precise and effective policy decision making leading to earlier detection and better prevention of emerging health threats than is currently the case.

Introduction

The use of secondary data represents an enormous potential for epidemic intelligence and research. Secondary data are defined as data being used for a purpose that differs from the intention for which the data were collected.[1] Typically, these data cover various application scenarios and come from a wide variety of pre-existing sources.[2] Data are understood as health-related if they provide information on health statuses of given individuals or populations. By using machine learning and cloud computing, secondary health-related data can substantially improve the detection and surveillance of emerging diseases.[3] In addition, research based on such applications potentially provides new insights into causes and consequences of diseases.[4]

Available sources of secondary health data differ in the degree to which they are used. For example, real-world data sources such as epidemic surveillance, disease registries, and population surveys as well as insurance, census, and government[5] records are widely used. Improvements in epidemic intelligence and research could be achieved by optimising the processing of data from such sources. However, the greatest potential for further progress arises from building comprehensive data ecosystems that additionally enclose data from sources presently on the verge of widespread use. These sources contain those with user-initiated content from data spaces that include the internet, wearables, and mobile phone apps,[6] and those with non-user-initiated content, such as electronic health records[7] and genome sequencing.[8] Figure 1 depicts real-world sources of a possible ecosystem for secondary health data.

The need to create data ecosystems for knowledge generation has been emphasised by WHO.[9] The continuous exchange of secondary health data within such systems plays a crucial role in strengthening policy and research activities aimed at achieving health equity for all. At the centre of these activities, which can be grouped under the heading of global health,[10] are epidemic intelligence and research in low-income and middle-income countries (LMICs). Strengthening epidemic intelligence and research by building data ecosystems requires wide-ranging multidisciplinary efforts. Depending on the objectives, strategies for building data ecosystems must be adapted to the given political environments, the available technical infrastructures, and the human capital.

This Health Policy paper intends to support the establishment of such strategies. The COVID-19 pandemic has highlighted the challenges of using secondary health data effectively. We present the results of an open exchange of basic considerations between key German research institutions and WHO on how ecosystems for an effective use of health data could be implemented. We first introduce data sources on the verge of widespread use and outline specific challenges and solutions in leveraging their related data. Reusability criteria as well as regulatory and ethical aspects are discussed in the subsequent section. We conclude with an outlook on the efforts still required.

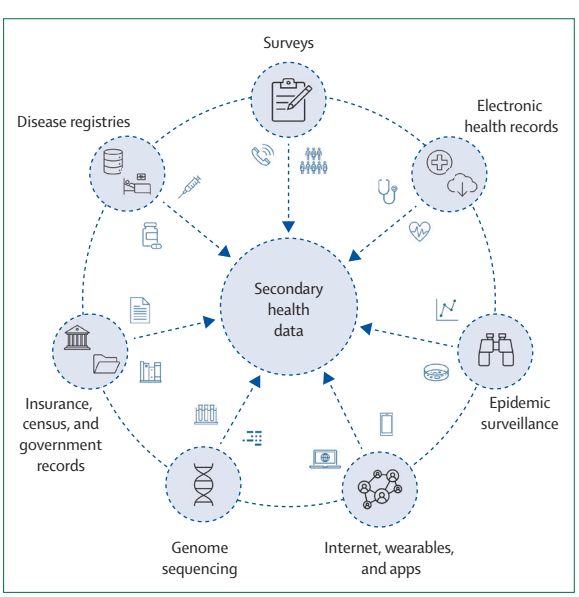

Figure 1: **Real-world sources of a possible ecosystem for secondary health data**

Lancet Digit Health 2023; 5: e93–101

Institute of Medical Informatics (A-F Näher MD, S A I Klopfenstein MD, Prof F Balzer PhD) **and Core Facility Digital Medicine and Interoperability, Berlin Institute of Health** (C N Vorisek MD, S A I Klopfenstein, M Lehne PhD, Prof S Thun MD), **Charité–Universitätsmedizin Berlin, Berlin, Germany; Method Development, Research Infrastructure, and Information Technology** (A-F Näher, L Grabenhenrich MD, Prof M von Kleist PhD), **Robert Koch Institute, Berlin, Germany** (Prof L H Wieler PhD); **Department of Mathematics and Computer Science, Free University Berlin, Berlin, Germany** (Prof M von Kleist); **Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia** (S Alsalamah PhD); **Digital Health and Innovation Department, Science Division, World Health Organization, Geneva, Switzerland** (S Alsalamah, S Pujari MSc); **Department of Data Science in Biomedicine, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg, Germany** (Prof D Heider PhD); **Department of Epidemiological Methods and Etiological Research** (Prof W Ahrens PhD) **and Department of Biometry and Data Management** (Prof I Pigeot PhD), **Leibniz Institute for Prevention Research and Epidemiology—BIPS, Bremen, Germany; Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany** (Prof W Ahrens, Prof I Pigeot); **Institute of Ethics, History, and Theory of Medicine, Ludwig Maximilian University of Munich, Munich, Germany** (Prof G Marckmann MD); **Institute for Planetary Health Behaviour, Department of**

Media and Communication Science, University of Erfurt, Erfurt, Germany (M A Jenny PhD)**; Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany** (M A Jenny)**; Harding Center for Risk Literacy** (M A Jenny) **and Digital Engineering Faculty, Hasso Plattner Institute** (Prof B Y Renard PhD)**, University of Potsdam, Potsdam, Germany; Health Communication Research Group (Implementation Science), Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany** (M A Jenny)

Correspondence to: Dr Anatol-Fiete Näher, Institute of Medical Informatics, Charité–Universitätsmedizin Berlin, Berlin 10115, Germany **anatol-fiete.naeher@charite.de**

Health data sources on the verge of widespread use

The internet, wearables, and mobile phone apps

New health-related data sources have become available with the internet, wearables, and mobile phone apps throughout the past decades. These sources' core strength is that they only rarely exhibit data flow discontinuities. Instead, they efficiently deliver information in near real time and at a high spatial resolution. Most web-based secondary health data are generated through health-related behaviour on social media and networks or search engine queries.[6] Entries in news media can also represent web-based data sources.

A great variety of different wearable devices and apps generate large amounts of personalised health data. Depending on the app, such data provide information ranging from distinct physiological parameters to more general medical conditions. The Corona Data Donation app released by the Robert Koch Institute, Germany's leading public health authority, is a prominent example of the use of data generated by wearables. To gain better epidemiological insights into the progression of the COVID-19 pandemic up to December, 2022, the software connected to wearable devices such as Apple's smartwatch, the FitBit, or the Oura ring and collected information on resting heart rate, body temperature, blood pressure, or physical activity and sleep.[11]

Symptom checkers such as Ada or Your.MD aim at providing users with insights into their individual health status by assessing symptoms and suggesting possible diagnoses. Although not relying on exact physiological measurements, the data generated by these apps are potentially of high importance for epidemic intelligence and research. For instance, Menni and colleagues[12] predicted positive results from SARS-CoV-2 PCR with information on self-reported symptoms obtained from a symptom checker.

Electronic health records

Assuming sufficient integration into clinical workflows, electronic health records can serve as a data source for epidemic intelligence and research. The main advantage of these records is that they can potentially make complete patient data easily accessible. Originally they were intended for billing and administrative purposes[7] and have later been adapted to better inform clinical decisions and reduce medical errors.[13] The most valuable information contained in electronic health records typically arises during routine care delivery such as physician notes and clinical imaging, or even demographic data. Recommendations on the meaningful use of electronic health records also suggest the inclusion of socioeconomic measures to assess health gradients more precisely.[14] In addition, distributed ledger technologies allow for electronic health record solutions whereby patient data can even be internationally registered across jurisdictions.[15]

Although more traditional surveillance frameworks typically rely on either syndrome definitions or laboratory results,[16] a key benefit of electronic health records-based epidemic intelligence and research exists in joining information from a flexible range of different domains. For instance, the tuberculosis surveillance module of the US Electronic medical record Support for Public Health depends on case definitions compiled from clinical data on prescriptions, test results, and previous diagnoses of tuberculosis as indicated by International Classification of Disease codes.[16] Comparable algorithms are incorporated into the electronic medical record architecture for the discovery of candidates for HIV pre-exposure prophylaxis[17] and surveillance of influenza[18] or diabetes,[19] among others. However, these examples should not obscure the fact that there are considerable barriers to the widespread use of electronic health records. A more detailed overview of the key organisational, technical, and human factors crucial to the successful adoption of electronic health records is provided by Fenelly and colleagues.[20]

Genome sequencing

In the past 10 years, the potential of molecular data on genomes has gained increasing attention in the fields of epidemic intelligence and research.[8] Although such information allows for tailored diagnostic and therapeutic approaches in clinical medicine, genetic data also help define population screenings and policy interventions more accurately. Similar works show how data generated by genome sequencing enable the tracking of spreading patterns of specific SARS-CoV-2 variants[21] or improve COVID-19 incidence estimates.[22] Whole-genome sequencing of extensively drug-resistant tuberculosis also facilitates the exact geographical location of such strains.[23] Further examples of genome data use for population health include recency estimations of HIV and hepatitis C viral infections[24,25] and outbreak detection of foodborne diseases.[26] Quality controlled genetic data are publicly available from various repositories such as GenBank,[27] the Sequence Read Archive,[28] GISAID,[29] and the European Nucleotide Archive.[30]

Barriers and possible solutions

Quality of secondary data

As well as the advantages of sourcing health-related data from the web, wearables, apps, or electronic health records there are also considerable limitations (table):[32,42] estimates solely resting on data gathered from these sources are prone to sampling bias. Depending on cultural and economic context, some user communities or patients with electronic health records are more or less likely to be selected into samples than other population members. Results obtained based on such samples are therefore not transferable to broader populations. Furthermore, correlations of data generated by internet, wearable, or app usage and parameters of interest might

	Type of analysis	Correction
Sampling bias, ie, bias resulting from unequal sample selection probabilities of observational units from a given population	Statistics	Weighing or raking
Sampling bias, ie, bias resulting from unequal sample selection probabilities of observational units from a given population	Machine learning	Data from differing domains can be combined such that accuracy is improved
Unobserved heterogeneity, ie, variation across observational units or time due to unobserved factors	Statistics	Fixed effects, difference-in-differences, regression discontinuity, instrumental variables, synthetic-control method, and machine learning methods for improving statistical estimates
Unobserved heterogeneity, ie, variation across observational units or time due to unobserved factors	Machine learning	Fixed effects or unit-specific time trends
Autocorrelation, ie, correlation of an observational unit with its time-lagged own values	Statistics and machine learning	Inclusion of error terms allowing for autocorrelation in models
Measurement error	Statistics	If the measurement error is: random, use instrumental variable methods for linear models (see Schennach[31] for non-linear specifications); systematic, see Schennach[31] for linear and non-linear specifications; causing misclassification of disease phenotypes, use estimators assuming either a bias depending on heterogeneity across patients,[32] or a joint time-dependent disease and disease-driven data generating process[33]
Measurement error	Machine learning	If the measurement error is: random or systematic, see Song and colleagues[34] for deep learning methods, Frénay and Verleysen[35] for non-deep learning methods; causing misclassification of disease phenotypes, there are no established methods yet
Missing data	Statistics and machine learning	In the case of: missing completely at random (also known as MCAR), observations can be deleted either listwise or pairwise; missing at random (also known as MAR), use imputation[36] or estimation by Full Information Maximum Likelihood or Expectation Maximation;[37] missing not at random (also known as MNAR), use estimators accounting for selection into non-missingness[38,39] or pattern mixture models[40,41]

Table: **Potential biases in secondary health data and means for correction**

vary over time and across individuals due to unobserved heterogeneity.[43] This type of bias could even occur in appropriate samples. Further considerations regarding the generating processes of web-based data are needed. These processes are very often affected by self-perpetuating patterns such as enfolding discussions on social media or news feeds. Models trained on such data are thus prone to autocorrelation bias, resulting in time-dependent variations in sensitivity and specificity.[44]

Validating internet, wearable, app, or electronic health record data with reference data from more conventional sources is a viable remedy to these shortcomings. By gauging the data-inherent biases, evaluation metrics can be constructed that enable the development of corresponding correction procedures. A readjustment of internet, wearable, app, and electronic health record samples is essentially possible by applying well known survey methods such as weighing or raking the data after collection.[32,44] In light of data collected with non-quantifiable sampling biases, an increasingly common solution exists in the combined analysis of data from different sources. Although respective procedures do not help obtain valid inference results, such approaches yield satisfying results concerning the prediction of disease incidences. For example, Quer and colleagues[45] augmented sensor data on resting heart rates and sleep and activity metrics with information from a symptom checker app. The authors were able to show that the combination of both data types discriminates significantly better between test results of individuals who are COVID-19 positive and individuals who are COVID-19 negative than models relying solely on the app data.

In the statistical analysis of observational secondary data, unit or time-varying unobserved heterogeneity can be controlled for by applying fixed effects, differences-in-differences, regression discontinuity, instrumental variable, or synthetic control designs. More recent approaches in this field also use machine learning methods to improve statistical estimates, especially when many covariates are present.[46] Unobserved heterogeneity only impairs the predictive performance of machine learning algorithms if correlations between observed features and unobservable features change over time. This circumstance can be accounted for by either eliminating any variance between observational dates or by including unit-specific time trends. Careful modelling choices are also essential for the correction of autocorrelation bias in internet-based data. By joining data from the Centers for Disease Control and Prevention with Google Trends' data, Yang and colleagues[47] showed that bias in Google Flu Trends resulting from alternating search behaviours of internet users can be captured by including error terms allowing for autocorrelation in machine learning models. Statistical models could be specified in the same manner.

Random or systematic measurement errors affect data quality, too. Random errors are the result of randomly distributed deviations in the measurement of a value

assumed to be true. Systematic errors occur due to non-random measurement deviations. Both error types lead to biased statistical estimates and impair machine learning predictions. In the case of electronic health records and app data, for instance, measurement errors might result in misclassification of disease phenotypes. The disease-related information obtained from such sources could be erroneous due to manifold causes. Insurance incentives possibly lead to preferential assignments of diagnostic codes included in electronic health records.[48] Specific symptom-related data included in electronic health records and apps could also be biased by over-reporting or under-reporting by patients. Furthermore, past patient histories are likely to be inconsistently recorded in electronic health records.[32]

Correction procedures must be specified according to the respective type of measurement error, the functional form of the specified model, and whether the error occurs in the dependent or independent variables. Instrumental variable methods can be used to account for random errors in linear statistical models. Schennach[31] reviews the available statistical approaches for non-linear random as well as linear and non-linear systematic measurement errors. For misclassification in electronic health records data due to systematic measurement error, there are two classes of non-linear statistical approaches for correction: Beesley and Mukherjee[32] presume that electronic health records misclassification is caused by factors varying across patients to derive estimators for misclassification rates. The second class of statistical models is suggested by Lange and colleagues.[33] As opposed to Beesley and Mukherjee,[32] the authors ground their method in the expectation of inter-individually constant sensitivity and specificity. Misclassification risks are then determined by taking a joint time-dependent disease and disease-driven observation process as a basis for the likelihood estimation of the observed electronic health records data.

Numerous robust machine learning methods are available for random and systematic measurement errors in training data. An overview of corresponding deep learning approaches is provided by Song and colleagues.[34] Frénay and Verleysen[35] present the existing non-deep learning methods. Machine learning methods that consider misclassification of disease phenotypes in electronic health records training data due to measurement error have not yet been developed.[49]

Missing data is another substantial source of bias. Existing strategies for bias correction need to be chosen with regard to distinct missingness patterns:[50] with data missing completely at random (also known as MCAR), observations with missing values could be deleted either listwise, ie, generally, or pairwise, ie, only in cases relevant to specific analyses. However, bias is likely to be induced by missing at random (also known as MAR) data—that is, if missingness depends on some observed trait but not on the missing data points themselves. A means for correction of bias owing to missing at random patterns is the imputation of missing values.[36] As a viable second option for statistics or machine learning, models can be fitted to missing at random data by Full Information Maximum Likelihood or Expectation Maximation.[37] A further bias emanates from missingness pertaining to certain values of one or more of the given data points. Solutions based on machine learning or statistical models for such missing not at random data (also known as MNAR) consist of specifications that account for selection into non-missingness.[38,39] An alternative is to partition the data and apply pattern mixture models that condition on subgroup-specific missing not at random patterns.[40,41]

Interoperability

To enable the exchange of health data within ecosystems, data need to be processed across countries and organisations. This approach requires the adoption of standard data formats and vocabularies providing human-readable and machine-readable data structures with unambiguous semantics. A means for accomplishing this goal is interoperability, ie, "the ability of two or more systems or components to exchange information and to use information that has been exchanged".[51] Interoperability is therefore not just defined by exchanging information but also by using the shared data in a meaningful way.

There are different levels of interoperability: technical interoperability provides basic data exchange capabilities between systems and syntactic interoperability specifies data format and structures.[52] The structured exchange of health data is supported by international standards development organisations. One major standard for communication of health data was established by Health Level Seven International (also known as HL7): Fast Healthcare Interoperability Resources (also known as FHIR). This standard defines data structures for typical health-care concepts, so-called resources, which can be modified and extended for specific use cases.[53] The advantage of Fast Healthcare Interoperability Resources is its reliance on existing internet technologies already supported by mobile devices. Furthermore, semantic interoperability uses terminologies, nomenclatures, and ontologies to define medical concepts unambiguously, making sharing them across worldwide systems possible. One of the largest terminologies to date is the Systematized Nomenclature of Medicine, Clinical Terms (also known as SNOMED CT). Additional domain-specific terminologies, such as the Logical Observation Identifiers Names and Codes (also known as LOINC) for laboratory data or the Human Phenotype Ontology (also known as HPO) for phenotypes, can complement the Systematized Nomenclature of Medicine, Clinical Terms.

Major steps need to be taken to promote data interoperability: policies should particularly aim for interoperable data exchange or even enforce interoperability through legal regulations.

Record linkage

Record linkage means the individual-level linkage of data records stored in separate databases to facilitate the combined analysis of person-related data from different sources. Such record linkage augments a dataset with additional information that was not originally included. An example is the assessment of incident cancer patients by linking cancer registries to data on prescribed medications collected by health insurers.

Linkage of individual-level records via person-identifying data such as name and address requires the informed consent of study participants. Such consent can usually be obtained in studies collecting primary data. Asking for informed consent is standard practice in studies such as the German National Cohort.[54] Here, privacy is preserved by pseudonymisation of study data through the agency of a trust centre, keeping person-identifying data and corresponding pseudonyms securely locked up.[55] However, if individual-level data from secondary data sources are to be linked, and informed consent is not feasible, legal restrictions often prohibit such linkage. Linkage of pseudonymised or anonymised datasets via individual characteristics might then be an alternative to linkage through person-identifying data. However, the necessary overlap in common variables allowing for this type of linkage with sufficient accuracy is rarely available in secondary health-related datasets.

Federated and swarm learning

Federated learning has been identified as a promising analytical tool for surveillance and research applications because it allows collaborators to keep their data private. This approach eliminates the need to share sensitive data between different parties:[56] instead, local machine learning models are trained at each partner site using local data. These models are then combined to improve overall predictive performance while maintaining a maximum amount of personal data protection. Going beyond federated learning, swarm learning is a recent technique maintaining confidentiality without the need for a central coordinator. As a decentralised machine learning approach, swarm learning unites edge computing, blockchain-based peer-to-peer networking, and coordination.[57] Figure 2 illustrates key differences in federated and swarm learning architectures.

The idea of federated learning can be divided into two types: horizontal and vertical federated learning. Horizontal federated learning is very close to the idea of ensemble and distributed learning. Different models are trained and aggregated in ensemble learning and horizontal federated learning, eg, by averaging. The training can be done in parallel, ie, all models are trained independently and then aggregated, or iteratively, ie, models are trained iteratively rather than independently. An example of iterative training is neural networks, typically trained in epochs, but these epochs are split among different parties in federated learning. The main difference between ensemble and federated learning is that the different models (and underlying data) come from different parties. In contrast to horizontal federated learning, vertical federated learning combines models with different features but identical samples. Federal techniques based on the same concepts as those presented previously are also available for fitting statistical models.

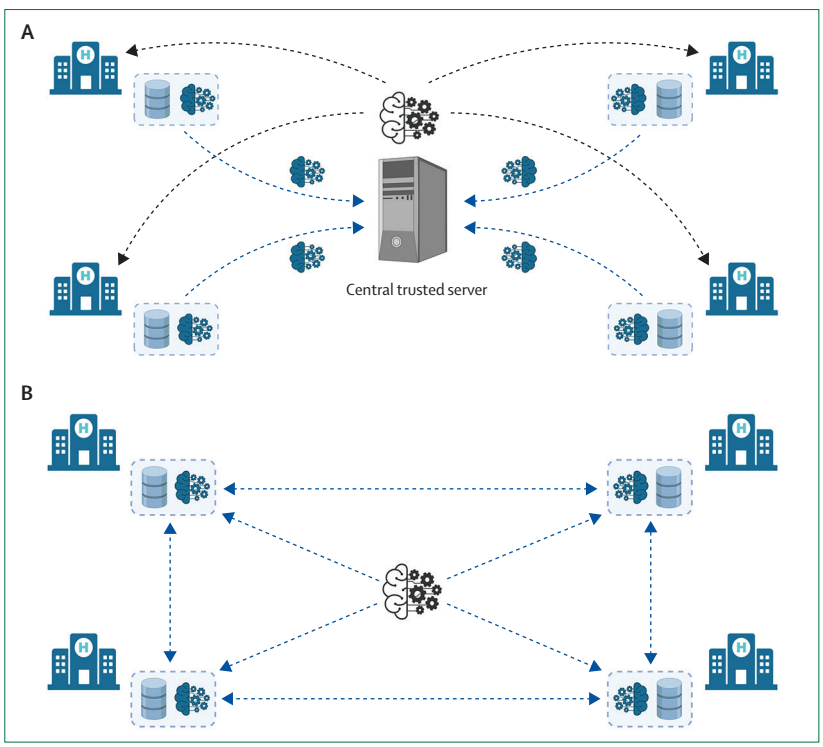

Figure 2: **Federated and swarm learning**
(A) The federated learning scheme includes a central trusted server used for communication between the partners and where the individual models trained at the different partner sites are combined. However, the data used for training the individual models never leave the safe information technology infrastructure at each partner site and are never disclosed. (B) The swarm learning scheme does not include a central server. Communication between the partners is carried out in a peer-to-peer manner, thus enabling an iterative improvement of local models. The data stay at the local information technology infrastructure at each partner site and are never disclosed. This figure was created with www.biorender.com.

A limitation is that federated learning requires different elements learned by respective partners to combine linearly. This restriction puts constraints on the plasticity of federated learning approaches to model real-life phenomena. Furthermore, implementing models in a federated manner does not inherently ensure privacy protection. Outliers can be detected in federated linear regression, and the underlying personal data can be back calculated through data and model poisoning, for example.[58] Implementers, therefore, need to weigh privacy requirements against the disclosure risks specific to different modelling strategies.

Data storage and harvesting

To continuously use and exchange secondary data, ensuring their availability is important. Besides

harvesting, data availability is achieved through storage in either data warehouses or data lakes. Data warehouses are centralised systems suitable for structured and consistent data.[59] Such data might originate from registries or surveys on health behaviours, for instance. Storage in warehouses requires extensive routines for data pre-processing. Data availability itself is maintained by comprehensive data models that contain meta-information of datasets and reference folder structures and file indices. Hence, maintenance and implementation of warehouse architectures are usually costly and time-consuming on the one hand. By their stable design, on the other hand, warehouses warrant high degrees of availability once the data have been entered.

Data lakes constitute a storage concept most suited for unstructured data. Such data are typically obtained from the internet, wearables, or apps. Within lakes, data are retained with their native formats.[60] Data schemes are thus not fixed until the data are being queried by users.[61] Unlike warehouses, data extraction depends on data models, including contextual layers describing the semantic meaning of data fields. Regarding implementation, the costs associated with such extensive metadata management need to be balanced against the advantages of lake systems that rapidly process large amounts of heterogeneous data.

In contrast to data warehouses and data lakes, harvesting systems index data objects by their sources and maintain data availability without storage. Harvesting requires metadata to reference all data in question. To build up metadata frameworks themselves, modern harvesting systems rely on routines for automated extraction of metainformation from multiple sources, such as text mining. Once indexed, data objects are linked via unique identifiers or stable links included in the metainformation. As identifiers or links need to be updated rather often, harvesting demands high maintenance efforts. Decision makers need to set the maintenance costs against the low implementation costs of harvesting systems that result from their lean architecture.

Key considerations for implementation

FAIR data

Any digitalisation initiative in global health faces the challenge of enabling the exchange of secondary data between countries and organisations for routine use. Such data come from manifold sources and are often incompatible with each other. To generally improve data management, the FAIR principles have been developed by a group of stakeholders from academia and industry.[62] The principles are thought of as general guidelines that ensure quality standards concerning the reusability of secondary data to develop data-driven applications.

FAIR data are defined as data that are findable, accessible, interoperable, and reusable. To be findable, data and metadata should be assigned a unique and persistent identifier and registered or indexed in a searchable resource. The principle of accessibility stipulates that data and metadata are retrievable by their identifier using a standardised, automated retrieval protocol. In addition, metadata should be accessible, even when the data are no longer available. To be interoperable, data and metadata should be expressed and shared using published standards for knowledge presentation. Moreover, reusability demands data and metadata to be described as clearly as possible with accurate and relevant attributes and clear and accessible conditions for use.

International regulations

Using secondary health data in accordance with FAIR principles remains extremely difficult due to the absence of international regulatory standards or guidance. The need to overcome this barrier has been acknowledged by the Focus Group on AI for Health.[63] The group has been set up by WHO and the International Telecommunication Union to build a framework for using artificial intelligence (AI) for health.[64] Concerning the regulatory concepts of AI for health, the Focus Group on AI for Health's Working Group on Regulatory Considerations identified six general topic areas: Documentation and Transparency, Risk Management and AI Systems Development Lifecycle Approaches, Intended Use and Analytical and Clinical Validation, Data Quality, Privacy and Data Protection, and Engagement and Collaboration. In an international, multidisciplinary effort involving regulatory bodies, policy makers, academia, and industry, the Working Group on Regulatory Considerations is working towards publishing an overview of the key regulatory concepts of AI technology deployment and AI systems development in health. The initiative is also preparing a list of relevant key recommendations for the way forward. Throughout the work of the working group, a particular emphasis is put on the introduction of international regulations for the exchange of secondary health data. Regarding such regulations, the working group intends to adopt data quality and protection regulations and standards that have been worked out in a benchmarking framework by the Focus Group on AI For Health. The framework incorporates task-specific gold standards for data modalities, annotations, and interoperability.[65]

Ethical considerations

Even assuming that FAIR data are available in accordance with international regulations, their use raises several ethical issues that must be adequately addressed. In addition to issues of safety and protection of sensitive information, the use of secondary data could have ambivalent implications on health equity. On the one hand, they could improve access to better disease prevention and health care in disadvantaged populations. On the other hand, some populations are partly or entirely excluded from health data coverage due to scarce

conventional or digital resources.[66] This shortfall potentially results in a digital divide regarding data-driven innovations for epidemic intelligence and research.[67] The main barriers affecting health data equity have been identified in a scoping review.[68] Using the Data Equity for Health and Health Equity framework based on these findings, O'Neil and colleagues[68] provide guidance on promoting equity in secondary health data. Access to data needs to be largely improved across communities and health-care sectors in LMICs. In line with FAIR data principles, improving data access means establishing open data policies and measures to strengthen technical, syntactic, and semantic interoperability. In addition, further action is needed to increase the usability of health data in LMICs. To better enable decision makers to gain insights, data should be provided in a ready-to-use format and the timespan for data dissemination should be shortened. According to the authors,[68] the availability of equitable secondary data in LMICs also depends on the quality of data collection. Quality can be ensured through standardisation of data collection methods and the training of people conducting the surveys. Moreover, health equity is largely enhanced by secondary data that include information about individuals' subjective perceived health states in addition to objective health outcomes. Subjectively perceived health states can be reliably recorded using questionnaires and compared using patient-reported outcome measures. The main value of these questionnaires is that, besides enhancing health equity, they contribute to an overall improvement in the quality of prevention and care.[69]

Closely related to data fairness are specific challenges in the context of machine learning applications to health data.[70] When trained on incomplete or skewed datasets, some algorithms have been shown to aggravate social inequalities in out-of-sample predictions.[71] But even unbiased training data could lead to machine learning models producing ethically questionable results if used in inappropriate environments. The transferability of the results of newly developed machine learning applications should therefore be tested on standardised datasets such as MIMIC-IV. MIMIC-IV is an open-access dataset that contains case-based information on clinical, laboratory, and administrative data from emergency departments and intensive care units at a large US hospital.[72] However, not all findings obtained from such datasets are representative of other health-care systems.

The root causes of questionable results can be particularly hard to identify when generated by black-box machine learning models.[73] International legal frameworks, such as the General Data Protection Regulation or the European Charter of Patients' Rights, thus require a minimum of explainability for machine learning-supported health-related decisions. While developing machine learning algorithms for epidemic intelligence and research, explainable models should be tested whenever possible. In this regard, the WHO guidance on Ethics and Governance of Artificial Intelligence for Health has been published as part of the Focus Group on AI for Health framework with contributions from leading experts in ethics, digital technology, law, and human rights, and experts from ministries of health.[74]

Ethical implications vary considerably and so each application involving the use of secondary health data needs to be assessed individually. Ethical assessments should follow a systematic approach on the basis of a predefined list of normative criteria.[75] The goal should be to identify ethically relevant implications and then provide guidance for the ethically justified design and implementation of digital tools in global health. In addition, individuals affected by data-based policy decisions should be informed about what data were collected, what was analysed about them, and how decisions were made.

Outlook

Harnessing the full potential of secondary data enables ground-breaking innovations in epidemic intelligence and research and sets new standards for decision makers. To make progress in this regard, the usability of secondary data needs to be substantially improved following FAIR and ethical principles. Most notably, policy makers should further strengthen their efforts to promote health data equity across sectors and global communities. The technical approaches and considerations we set out are not equally important for each of

Search strategy and selection criteria

This Health Policy paper is based on the results of an open exchange between German research institutions and WHO. The exchange was initiated with the goal of enabling future use of secondary health data that would help strengthen surveillance and research in a sustainable manner. It was motivated by the inadequacies of existing infrastructures for secondary use of health data that became apparent during the COVID-19 pandemic. Each of the institutions brought their specific perspectives and expertise to first outline the technical framework of the paper. This outline was accomplished by identifying topic areas deemed relevant to building comprehensive ecosystems for secondary health data exchange. On the basis of the identified topic areas, the following search terms for the literature review were derived by consensus of all authors: "global health", "public health surveillance", "public health internet", "genetics public health", "electronic health records", "causality", "measurement error", "label noise", "missing data", "selection bias", "interoperability", "HL7", "record linkage", "federated learning", "data lake", "FAIR principles", "data equity", and "patient reported outcome measures". Using these terms, the relevant literature was identified by the authors primarily responsible for each of the manuscript's topic-related sections. English language was imposed as a restriction on all searches. Papers from searches were included if they presented technical approaches or theoretical frameworks considered relevant for the processing and analysis of secondary health data for global health digitalisation. In addition, to avoid overlap, only papers that represented an extension of the content of the literature already identified were considered. Initial literature searches were conducted between April 1 and Dec 31, 2021, via Google Scholar, yielding 31 papers. A second round of searches took place between June 20 and July 12, 2022, which led to the identification of another 8 papers. In addition to these, 36 papers were selected based on the expertise of the respective authors.

the specific problems involved. Instead, relevant aspects enable the use of secondary health data in one of the numerous application areas, such as short-term detection of disease outbreaks, tracking of transmission patterns, accurate epidemic forecasting, or precise assessments of envisaged non-pharmacological interventions.

In general, the goal must be to share health-related data and insights for the common good. This goal is made possible through broad engagement and collaboration between different disciplines and stakeholders. Thereby, substantial progress can be made in global health.

Contributors

A-FN conceptualised, edited, and critically reviewed the manuscript. He prepared the introduction and section on health data sources on the verge of widespread use and conceptualised and contributed to the preparation of the subsection on quality of secondary data and prepared the data storage and harvesting subsection. FB conceptualised the introduction, contributed to the conceptualisation of the subsection on electronic health records, conceptualised and contributed to the preparation of the subsection on quality of secondary data, and critically reviewed the manuscript. MvK conceptualised and critically reviewed the introduction and section on health data sources on the verge of widespread use. LHW conceptualised the subsection on the internet, wearables, and apps and critically reviewed the manuscript. ML and ST conceptualised and prepared the subsection on interoperability. IP and WA conceptualised and prepared the subsection on record linkage. DH conceptualised and prepared the subsection on federated and swarm learning. CNV and SAIK conceptualised and prepared the subsection on FAIR data and critically reviewed the subsections on electronic health records and interoperability. SA and SP conceptualised and prepared the subsection on international regulations. GM and MAJ conceptualised and prepared the subsection on ethical considerations. BYR critically reviewed the manuscript and contributed to the conceptualisation of the subsection on electronic health records and conceptualised the genome sequencing subsection and outlook section. LG critically reviewed the manuscript and prepared the outlook section. The final version of the manuscript has been reviewed and approved by all authors.

Declaration of interests

We declare no competing interests.

Acknowledgments

This work has been made possible by the funding received as part of the NFDI4Health project financed by Deutsche Forschungsgemeinschaft (grant number 442326535). MvK also acknowledges funding from the Germany Ministry for Science and Education (grant number 01KI2016).

References

1 Vogt WP, Johnson B. Dictionary of statistics & methodology: a nontechnical guide for the social sciences. New York, NY: Sage, 2011.
2 Angrist JD, Krueger AB. Empirical strategies in labor economics. In: Ashenfelter O, Card D, eds. Handbook of labor economics, vol 3. Amsterdam: Elsevier, 1999: 1277–366.
3 Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. *Milbank Q* 2014; **92:** 7–33.
4 Eysenbach G. 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* 2009; **11:** e11.
5 Perez JAR, Emilsson C, Ubaldi B. OECD Open, Useful, and Re-usable data (OURdata) Index: 2019. Organisation for Economic Co-operation and Development Policy Papers on Public Governance, 2020.
6 Althouse BM, Scarpino SV, Meyers LA, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Sci* 2015; **4:** 1–8.
7 Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. *Annu Rev Public Health* 2016; **37:** 61–81.
8 Brand A, Brand H, Schulte in den Bäumen T. The impact of genetics and genomics on public health. *Eur J Hum Genet* 2008; **16:** 5–13.
9 WHO. WHO hub for pandemic and epidemic intelligence. Better data. Better analytics. Better decisions. Geneva: World Health Organization, 2021.
10 Koplan JP, Bond TC, Merson MH, et al. Towards a common definition of global health. *Lancet* 2009; **373:** 1993–95.
11 Corona-Datenspende. Our reports. 2022. https://corona-datenspende.de/science/en/ (accessed Dec 26, 2022).
12 Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med* 2020; **26:** 1037–40.
13 Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med* 2010; **363:** 501–04.
14 Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. Capturing social and behavioral domains and measures in electronic health records: phase 2. Washington DC: National Academies Press, 2014.
15 Alsalamah S, Alsalamah HA, Nouh T, Alsalamah SA. HealthyBlockchain for global patients. *Comput Mater Continua* 2021; **68:** 2431–49.
16 Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health* 2015; **36:** 345–59.
17 Krakower DS, Gruber S, Hsu K, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. *Lancet HIV* 2019; **6:** e696–704.
18 Yih WK, Cocoros NM, Crockett M, et al. Automated influenza-like illness reporting—an efficient adjunct to traditional sentinel surveillance. *Public Health Rep* 2014; **129:** 55–63.
19 Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care* 2013; **36:** 914–21.
20 Fennelly O, Cunningham C, Grogan L, et al. Successfully implementing a national electronic health record: a rapid umbrella review. *Int J Med Inform* 2020; **144:** 104281.
21 Hodcroft EB, Zuber M, Nadeau S, et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* 2021; **595:** 707–12.
22 Smith MR, Trofimova M, Weber A, Duport Y, Kühnert D, von Kleist M. Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. *Nat Commun* 2021; **12:** 6009.
23 Shah NS, Auld SC, Brust JC, et al. Transmission of extensively drug-resistant tuberculosis in South Africa. *N Engl J Med* 2017; **376:** 243–53.
24 Carlisle LA, Turk T, Kusejko K, et al. Viral diversity based on next-generation sequencing of HIV-1 provides precise estimates of infection recency and time since infection. *J Infect Dis* 2019; **220:** 254–65.
25 Carlisle LA, Turk T, Metzner KJ, et al. HCV genetic diversity can be used to infer infection recency and time since infection. *Viruses* 2020; **12:** 1241.
26 Lachmann R, Halbedel S, Lüth S, et al. Invasive listeriosis outbreaks and salmon products: a genomic, epidemiological study. *Emerg Microbes Infect* 2022; **11:** 1308–15.
27 Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res* 2013; **41:** D36–42.
28 Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nucleic Acids Res* 2011; **39:** D19–21.
29 Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data—from vision to reality. *Euro Surveill* 2017; **22:** 30494.
30 Harrison PW, Ahamed A, Aslam R, et al. The European Nucleotide Archive in 2020. *Nucleic Acids Res* 2021; **49:** D82–85.
31 Schennach SM. Recent advances in the measurement error literature. *Annu Rev Econ* 2016; **8:** 341–77.

32 Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics* 2020; **78:** 214–26.

33 Lange JM, Hubbard RA, Inoue LYT, Minin VN. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics* 2015; **71:** 90–101.

34 Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. *IEEE Trans Neural Netw Learn Syst* 2022; published online March 7. https://doi.org/10.1109/TNNLS.2022.3152527.

35 Frénay B, Verleysen M. Classification in the presence of label noise: a survey. *IEEE Trans Neural Netw Learn Syst* 2014; **25:** 845–69.

36 van Buuren S. Flexible imputation of missing data, 2nd edn. New York, NY: CRC, 2018.

37 Enders CK. Applied missing data analysis. New York, NY: Guilford Press, 2010.

38 Heckman JJ. Sample selection bias as a specification error. *Econometrica* 1979; **47:** 153–61.

39 Cortes C, Mohri M, Riley M, Rostamizadeh A. Sample selection bias correction theory. In: Freund Y, Györfi L, Turán G, Zeugmann T, eds. Algorithmic learning theory. Berlin and Heidelberg: Springer, 2008: 38–53.

40 Little RJ. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc* 1993; **88:** 125–34.

41 Ghalebikesabi S, Cornish R, Holmes C, Kelly L. Deep generative missingness pattern-set mixture models. 2021. https://proceedings.mlr.press/v130/ghalebikesabi21a/ghalebikesabi21a.pdf (accessed Nov 30, 2021).

42 Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. *Science* 2014; **343:** 1203–05.

43 Choi J, Cho Y, Shim E, Woo H. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health* 2016; **16:** 1238.

44 Aiello AE, Renson A, Zivich PN. Social media–and internet-based disease surveillance for public health. *Annu Rev Public Health* 2020; **41:** 101–18.

45 Quer G, Radin JM, Gadaleta M, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nat Med* 2021; **27:** 73–77.

46 Athey S, Imbens GW. The state of applied econometrics: causality and policy evaluation. *J Econ Perspect* 2017; **31:** 3–32.

47 Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc Natl Acad Sci USA* 2015; **112:** 14473–78.

48 O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005; **40:** 1620–39.

49 Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; **178:** 1544–47.

50 Little RJ, Rubin DB. Statistical analysis with missing data, vol 793. Hoboken: John Wiley & Sons, 2019.

51 Geraci A. IEEE standard computer dictionary: compilation of IEEE standard computer glossaries. Piscataway, NY: Wiley-IEEE Press, 1991.

52 Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. *NPJ Digit Med* 2019; **2:** 79.

53 Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. 2013. http://www.cs.ecu.edu/sartipi/papers/CBMS2013.pdf (accessed May 12, 2021).

54 German National Cohort (GNC) Consortium. The German National Cohort: aims, study design, and organization. *Eur J Epidemiol* 2014; **29:** 371–82.

55 Stallmann C, Ahrens W, Kaaks R, Pigeot I, Swart E, Jacobs S. Individual linkage of primary data with secondary and registry data within large cohort studies—capabilities and procedural proposals. *Gesundheitswesen* 2015; **77:** e37–42 (in German).

56 Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol* 2019; **10:** 1–19.

57 Warnat-Herresthal S, Schultze H, Shastry KL, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* 2021; **594:** 265–70.

58 Bhagoji AN, Chakraborty S, Mittal P, Calo S. Analyzing federated learning through an adversarial lens. 2019. https://proceedings.mlr.press/v97/bhagoji19a.html (accessed Sept 9, 2021).

59 Jarke M, Quix C. On warehouses, lakes, and spaces: the changing role of conceptual modeling for data integration. In: Cabot J, Gómez C, Pastor O, Sancho MR, Teniente E, eds. Conceptual modeling perspectives. Cham: Springer, 2017: 231–45.

60 Miloslavskaya N, Tolstoy A. Big data, fast data, and data lake concepts. *Procedia Comput Sci* 2016; **88:** 300–05.

61 Sawadogo P, Darmont J. On data lake architectures and metadata management. *J Intell Inf Syst* 2021; **56:** 97–120.

62 Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; **3:** 160018.

63 International Telecommunication Union. Focus Group on "Artificial Intelligence for Health". https://www.itu.int/en/ITU-T/focusgroups/ai4h/pages/default.aspx (accessed Jan 1, 2023).

64 Wiegand T, Krishnamurthy R, Kuglitsch M, et al. WHO and ITU establish benchmarking process for artificial intelligence in health. *Lancet* 2019; **394:** 9–11.

65 Wiegand T, Lee N, Pujari S, et al. Whitepaper for the ITU/WHO Focus Group on Artificial Intelligence for Health. 2020. https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf (accessed Dec 20, 2021).

66 Makri A. Bridging the digital divide in health care. *Lancet Digit Health* 2019; **1:** e204–05.

67 McAuley A. Digital health interventions: widening access or widening inequalities? *Public Health* 2014; **128:** 1118–20.

68 O'Neil S, Taylor S, Sivasankaran A. Data equity to advance health and health equity in low- and middle-income countries: a scoping review. *Digit Health* 2021; published online Dec 22. https://doi.org/10.1177/20552076211061922.

69 Black N. Patient reported outcome measures could help transform healthcare. *BMJ* 2013; **346:** f167.

70 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; **25:** 1337–40.

71 Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. *Nature* 2018; **559:** 324–26.

72 Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.0). *Circulation* 2022; **101:** e215–20.

73 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018; **15:** e1002683.

74 WHO. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization, 2021.

75 Marckmann G. Ethical implications of digital public health. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2020; **63:** 199–205 (in German).