The Journal of Supercomputing (2023) 79:19574–19606
https://doi.org/10.1007/s11227-023-05410-0

Data quality model for assessing public COVID-19 big datasets

Alladoumbaye Ngueilbaye[1,2] · Joshua Zhexue Huang[1,2] · Mehak Khan[3] · Hongzhi Wang[4]

Accepted: 17 May 2023 / Published online: 31 May 2023

Abstract

For decision-making support and evidence based on healthcare, high quality data are crucial, particularly if the emphasized knowledge is lacking. For public health practitioners and researchers, the reporting of COVID-19 data need to be accurate and easily available. Each nation has a system in place for reporting COVID-19 data, albeit these systems' efficacy has not been thoroughly evaluated. However, the current COVID-19 pandemic has shown widespread flaws in data quality. We propose a data quality model (canonical data model, four adequacy levels, and Benford's law) to assess the quality issue of COVID-19 data reporting carried out by the World Health Organization (WHO) in the six Central African Economic and Monitory Community (CEMAC) region countries between March 6,2020, and June 22, 2022, and suggest potential solutions. These levels of data quality sufficiency can be interpreted as dependability indicators and sufficiency of Big Dataset inspection. This model effectively identified the quality of the entry data for big dataset analytics. The future development of this model requires scholars and institutions from all sectors to deepen their understanding of its core concepts, improve integration with other data processing technologies, and broaden the scope of its applications.

Keywords Data quality model · COVID-19 big dataset · 4A · Canonical data model · Benford's law · CEMAC region

1 Introduction

The fast spread and magnetic nature of the COVID-19 pandemic have emphasized the importance of data quality, analyses, and models describing the potential trajectory of COVID-19 to comprehend its effects on the planet. The Centres for Disease Control and Prevention (CDC) and the World Health Organization (WHO) have issued a set of general recommendations and technical guidelines as the frontrunners

Extended author information available on the last page of the article in the fight against the new coronavirus [1, 2]. In the Central African Economic and Monitory Community(CEMAC) region countries, the CDC used rudimentary surveillance systems to gather COVID-19 data in collaboration with provinces, local, and other partners [3]. Each country has its reporting system, such as dashboards, daily bulletins, etc. The reported COVID-19 data through these methods varies greatly from one country to the other in terms of both content and format. In order to properly analyze the data produced by these surveillance technologies, decision-makers and analysts must be aware of their limitations. For instance, surveillance data on the number of reported cases of COVID-19 is inaccurate for several reasons, and they do not reflect the actual number of cases [4]. For effective epidemic control, reliable epidemiological surveillance systems are required. One of its main purposes is to offer reliable data quality to make informed judgements [5]. In relation to the COVID-19 pandemic, recent studies [5–7] assessed the reliability of COVID-19 data gathered by the World Health Organization (WHO), the European Centre for Disease Prevention and Control (ECDPC), and the Chinese Centre for Disease and Prevention (CCDP). This study discovered that measurement inaccuracies increased and became more obvious when more countries supplied data to the official repositories [5]. It is challenging to understand the quality of the dataset during the COVID-19 pandemic.

Benford's law was developed as a simple and rapid technique to evaluate the effectiveness of surveillance systems during the influenza A (H1N1) pandemic. Its significance has been further highlighted by the Zika pandemic in American countries and the dengue outbreak in Paraguay from 2009 to 2011 [8]. Governments and public health systems require precise and timely information about the characteristics and behavior of COVID-19 data in order to effectively address this ongoing public health crisis. When the COVID-19 epidemic started, several African nations quickly expanded their diagnostic and surveillance capabilities [9]. The early implementation of non-pharmaceutical treatments to stop transmission in the majority of African nations came at a cost to their healthcare systems and post-disaster rehabilitation [10, 11]. Data reporting techniques vary widely because of the diversity between African nations and the range of COVID-19 experiences. The COVID-19 cases, deaths, and recoveries reported by recognized regional cooperating centers and member nations are compiled by the Africa Centre for Disease Control and Prevention (ACDC) using an online dashboard [12]. However, nations also have national reporting programs that can give researchers, medical professionals, policymakers, and the general public access to more thorough data. The frequency, substance, and style of such national reporting systems can vary and they may be more difficult to access.

Presently, it is practically challenging to produce high-quality statistics in this pandemic era without extensive cooperation/collaboration between health authorities, healthcare providers, and researchers from other sectors. Due to the COVID-19 pandemic's urgency, datasets of poorer quality may be used, endangering the validity of the conclusions and leading to biased evidence. Poor decision-making or not using data to inform decisions could be the results. Access to high quality health data is one of the methodological difficulties connected with evaluating COVID-19 data during the pandemic, and various data quality issues have been discussed [13–15]. However, to our knowledge, no existing study systematically evaluates the data quality problems in the datasets provided by the national surveillance systems of the six CEMAC region countries for research purposes during the COVID-19 pandemic. Despite this being a global issue, this study solely uses WHO datasets for the six countries in the (CEMAC) region.

This paper proposes data quality model (DQM) (see Fig. 1), which comprises of three main propositions: (1) we propose a canonical data model (CDM), which is a standard data model to construct a common query that yields common results (see Fig. 2). Uncommon disease or outcomes of millions of patients cannot be investigated by a single institution. Thus, there is a constant push to build massive data networks across organizations, regions, and countries. You might be in charge of setting up a sizable data network that connects various organizations, each of which has its own electronic health records (EHR) system (see Fig. 3). Even hospitals with software for medical records from the same supplier adapt it to their specific patient populations, business requirements, and clinical workflow. There are other reliable data sources than the electronic health records (EHR). Institutions may also keep data from internal warehouses, patient registries, and various other data architectures. (2) we propose four adequacy levels (4A) as data quality model to fulfill the goals and purposes of the analysis. When we began by identifying the main characteristics of the data quality for our model, accessibility, contextual, representation, and intrinsic classification were taken as a basis, and (3) Benford's law, which is a rapid tool to test the quality of the dataset for the different selected countries.

Contributions. In this paper, our contributions are stated as follows: The data quality models are seen to have the potential to significantly influence the decision-making process for data assessment. It can be an effective tool for the scientific management of outbreaks while giving decision-makers situational insights from other facilities. To this end, the main contributions of this paper in response to these aforementioned problems are the following.

1. Data quality model with four levels of adequacy (4A) has been suggested in order to comply with the unique features of the analysis that is carried out using a particular Big Data solution (see Table 2);
2. Canonical data model has been proposed and analyzed to design a common query that yields common results;
3. Enhanced Benford's law is implemented to test the quality of the collected datasets on various countries from the CEMAC region countries;
4. Experiments are performed on real datasets collected from March 6, 2020, and June 22, 2022.

This paper is structured as follows. The related work is presented in Sect. 2. Background and the key components of data quality are covered in Sect. 3. Our proposal is shown in Sect. 4. experimental analysis in Sect. 5. The benefits of the model in Sect. 6, and the limitations in Sect. 7. The paper is concluded with Sect. 8.

2 Related works

The existing data quality evaluation methods for COVID-19 data focus only on outbreak prediction and data quality evaluation [3, 4, 9, 16] in the COVID-19 surveillance systems data. Through the year 2020, Judson et al. [9] looked at national COVID-19 reporting procedures in all 54 African nations. These reporting systems were broken down into three categories: frequency, report type, and data substance. Healthcare capacity, diagnostic testing, and reporting parameters related to patient demographics and morbidities were compared. Ashofteh et al. [4] employed comparative statistical analysis to estimate the accuracy of data collected by the CCDCP, WHO, and ECDPC) based on the value of systematic measurement errors. The success of government actions against COVID-19 and the revelation of data quality issues that jeopardize the validity of non-pharmaceutical therapies were discussed in [16]. Further, Idrovo et al. [3] investigated on the success of COVID-19 pandemic monitored by the Chinese epidemiological monitoring system. They used information from the WHO's situation reports 1 through 55 to fill in the gaps in Benford's law in order to provide an answer to this query. They came to the conclusion that throughout the present health emergency, the Chinese epidemiological surveillance system had good data quality. We note, however, that these studies have only looked at a few COVID-19 data quality problems and their potential solution.

A study assessed the effectiveness of Tunisia's system for monitoring influenza-like illnesses, concluded that it needs to be closely monitored and improved to provide a more accurate picture of the disease situation [17]. In Nigeria, Visa et al. [18] examined the Kano State malaria surveillance and suggested methods to enhance the quality of the data. Reliable and timely data will help researchers, public health officials, and the general public assess how the coronavirus pandemic is affecting healthcare systems and prepare for the proper policy response at all levels of government [19]. Currently, the alternative mathematical models created for other diseases and/or the experience of other nations where the outbreak has been discovered early and developed are being used as the basis for decisions and actions by governments and policymakers worldwide. For public health, a data-intensive field, high-quality, institutionally based datasets are necessary for data management and analysis [20]. Effective data quality model would ensure congruent results from many studies conducted worldwide during the data gathering. In this paper, we propose data quality model to tackle COVID-19 data quality issues in the CEMAC region countries. This model is used to evaluate and suggest the COVID-19 data so that we can have a credible basis for decision-making. It can also be applied for other countries as well.

3 Background

This section discusses the perspective of COVID-19 data, issues of their collection, and the overview of data quality model framework measurement.

3.1 The perspective of COVID-19 data quality

Big data currently lacks a structural definition, making it challenging to determine the proper concept of data quality for big data. According to Gartner [21–23], big data is a significant, quick-changing, and extremely diverse information asset that calls for creative and economical information processing to enhance decision-making, visibility, and process automation. Loshin et al.,[24] asserted that, when resource requirements surpass the capabilities of the current data technology environment, big data are subject to cost-effective solutions to handle current and future business difficulties. Moreover, big data is a comprehensive term that includes the dataset itself, along with technical, medical, commercial, and spatial value issues [23, 25, 26]. Big data are useful primarily because it extracts significant business value from data. High management tends to a propensity to believe that the bigger the big data, the greater the potential benefits. Regrettably, this occurs even when they are unsure of precise handling of big data issues or how to derive possible knowledge from big data research [22, 27]. Therefore, the first step in any big data research is to support high management in leading the research rather than investing in and implementing sophisticated technologies that will not yield any results that are pertinent to the research issue at hand [25, 28, 29].

In general, data quality management is concerned with evaluating datasets and taking remedial measures to make sure that the datasets serve the original goals for which they were designed [30]. To put it another way, the incoming data are worthwhile and suitable for big data analysis. The implementation of data quality management principles differs slightly from the application of normal data due to additional technological and organisational obstacles brought about by big data [31]. Some of these facts are gathered in Table 1.

3.2 COVID-19 data quality issues

COVID-19 has been around since 2019, and since its inception, tremendous amounts of data have been collected and disseminated in real-time. These data are used to inform various decisions regarding the implementation of non-prescription interventions and public health policies. Despite the increasing number of terms used in big data, such as velocity, volume, and variety, they do not allow for the necessary quality checks to be performed properly. This is because the data quality issues that are typically associated with this type of data are becoming more evident. In addition to having a big volume, data quality also requires a great deal of effort to improve its application-level and scale. This is because the exploration of reliable and efficient data is very important to the development of new knowledge. Along with a large volume, big data also needs outstanding data quality. Because the investigation of effective and dependable data quality is crucial, it is necessary to achieve the simultaneous improvement of data scale, quality, and application-level.

In Africa, the CEMAC region countries registered its first cases in Cameroon on March 6, 2020, in Yaoundé. The second country to register his first was Gabon on

Table 1 Comparing the classification of data quality for large datasets to the traditional classification [22, 23, 26, 28]

Technical and management issues	Data quality and and typical data relationships	Big dataset quality
Confidence levels based on processing frequency and variety	Batch-oriented processing. Structured data are mainly used in data warehouses. Data must be in initial conditions for analysis	Treatment in real-time and batch prompts. Data types can be structured, disorganized (unstructured), and semi-structured. It is necessary to have sufficient data in order to clean noise. This may or may not affect analysis to collect information about companies with poor quality
Data cleansing time	In the data warehouse, data are cleaned before it is loaded	It may be difficult to read data as-is since it may not fully understand the important elements and their relationships. To reduce capacity requirements, analyze data size and memory stream speed
Data points of importance	Data quality is assessed for critical data items such as country code and total cases	It may be necessary to conduct advanced exploration of data that is close or poorly defined. Therefore, important data items may change frequently
Analytical area	Afterward, a data quality engine and an evaluation engine are used to evaluate the data	For acceptable processing speed, analysis engines and data quality can switch to the data
Policy	Gaining, storing, analyzing, and backing up information costs money	Long-term data retention requires a retention strategy approach

March 12, 2020, while countries in the third phase were Central Africa Republic, Congo Brazzaville, and Equatorial Guinea, which reported their case on the March 14, 2020. The last country was Chad which recorded its case on the March 19, 2020. However, most of the identified cases were imported which originated from Europe and the United States of America rather than China. Because many of the healthcare systems in the continent were weak and had issues such a lack of financing and equipment, inadequate training of healthcare professionals, and ineffective data transfer, experts were concerned that COVID-19 would expand to Africa. It was believed that the pandemic, if it spread significantly, would be difficult to contain in Africa and would seriously affect the continent's economy.

3.3 COVID-19 data quality model and measurements

The data quality model (DQM) outlines a number of theoretical underpinnings and principles for impartially and objectively assessing the quality of data. An important idea of the model and its relationships is shown in Fig. 5. As a result, it can be used to assess various quality measurement data's quality features. The quality attribute of the data is evaluated quantitatively by the quality measure specified in this model. The entire data life-cycle stage as well as additional operations can make use of data quality measurement, including:

- establishing criteria for data quality;
- evaluating data quality;
- implementing and maintaining data management, and processing documents;
- putting in place and maintaining the IT service management process;
- to improve the effectiveness and quality of data used in the organization's decision-making processes;
- during the investigation, to compare the data quality of several data management technologies;
- to evaluate the system's performance and/or the data-generation procedures.

4 Methodology

In this section, we give an extensive explanation of the operational procedure and the concept of the data quality model framework structured in Fig. 1, which is our framework where data flow through several modules to assess its quality. The primary modules of this framework consist of (a) data profiling with the four adequacies (4A) Model, (b) Canonical Data Model to uniform the CEMAC region countries systems, and (c) Data Quality Assessment through Benford's law. The large impact poor data has on analytics always justifies the capability of thorough analysis and evaluation of the quality of the big dataset. All organizations from different domains of expertise rely on data when planning short- or long-term strategies. Developing data quality models with poor data are of particular significance. In order to address this issue effectively, the model should be used to investigate an adequate solution.

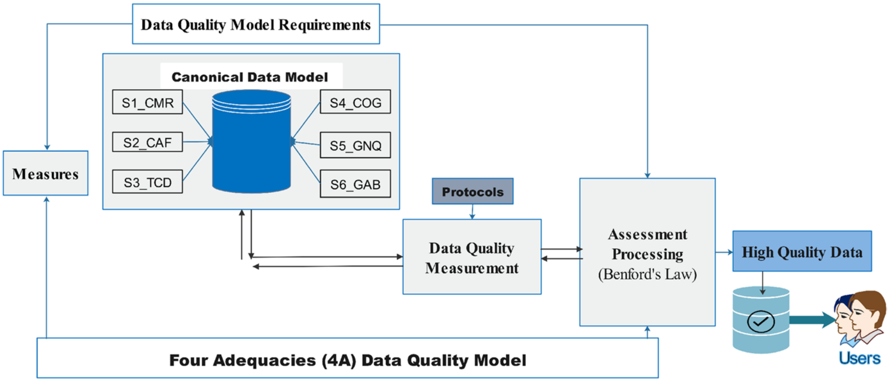

Fig. 1 Data quality model framework

Every model has strengths and weaknesses depending on the question posed on the ground.

4.1 The framework of the study

The Data Quality Model (DQM) framework is illustrated in Fig. 1. It is our theoretical framework where data flows through several modules to gauge its correctness. The main components of our model are (a) data profiling with the four adequacies (4A) Model, (b) canonical data model to uniform the CEMAC region countries systems, and (c) data quality assessment through Benford's law. This framework comprises three main structures: (1) the canonical data model (CDM), which is a standard data model to construct a common query that yields common results as shown in Fig. 2. Uncommon disease or outcomes of millions of patients cannot be investigated by a single institution. Thus, there is a constant push to build massive data networks across organizations, regions, and countries. One might be in charge of setting up a sizable data network that connects various organizations, each of which has its own electronic health records (EHR) system (see Fig. 3). (2) the four adequacy levels (4A) as data quality model to fulfill the goals and purposes of the analysis, and (3) Benford's law, which is a rapid tool to test the quality of the dataset for the different selected countries.

4.2 COVID-19 data quality model

Building the statistical and computational techniques for the data quality model approach, we firstly propose the mathematical framework to illustrate the COVID-19 concept. Secondly, we propose the four adequacies as a data quality model. Thirdly, we propose common data model also known as the CDM, to enable CEMAC region countries to establish and distribute a common definition of its entire data unit. Lastly, we propose Benford's law to assess the datasets' quality. To improve decision-making, the data quality model seeks to address data quality before data

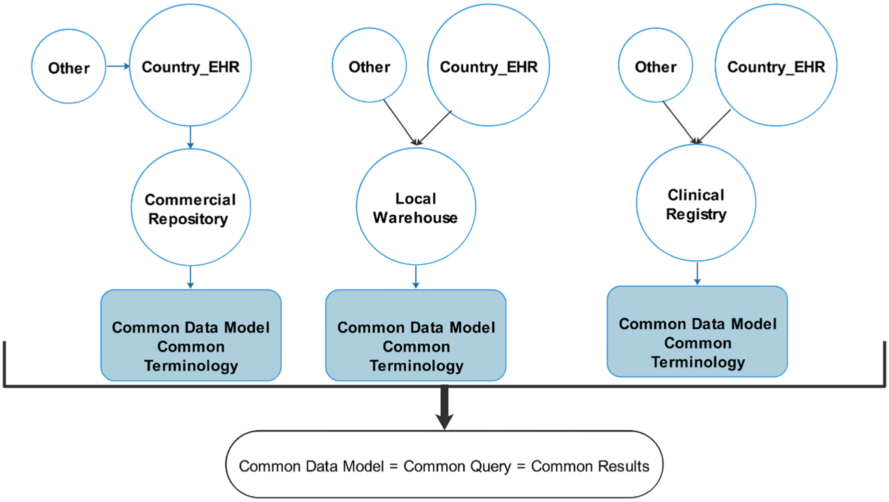
Other
Country_EHR
Other
Country_EHR
Other
Country_EHR
Commercial Repository
Local Warehouse
Clinical Registry
Common Data Model Common Terminology
Common Data Model Common Terminology
Common Data Model Common Terminology
Common Data Model = Common Query = Common Results

Fig. 2 Data model with common results

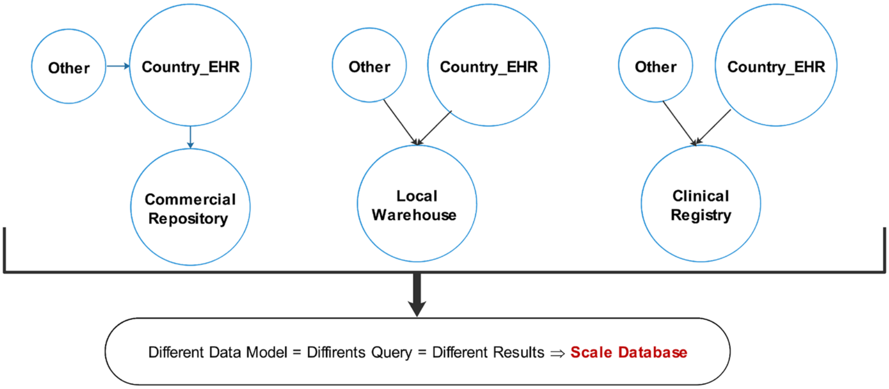

Fig. 3 Data model with different results

analytic. This is accomplished by estimating the properties or features of the data's quality and using the data quality model metric to assess the attributes' accuracy, completeness, and consistency. Moreover, the benefits of this common data model are to increase the ability to include more data partners to participate in the data request; it reduces the need for database administrators to reformat the output, and it ensures that the same query logic is applied at all data partners. This research mainly deals with data quality of data sources, more precisely COVID-19 datasets of CEMAC region countries. The assessment is essential to ensure the quality levels of these countries' COVID-19 datasets with optimal costs. Hence, we must underline the importance of big dataset quality since without it, we are unable to generate accurate estimations for the analytic.

4.2.1 Formulation

Let $\mathbb{D}$ represents a set of six countries' datasets with m instances and n observations, i.e., D has m data samples. Each sample has n attributes. $\mathbb{D}$=\{$S1\_CMR$, $S2\_CAF$, $S3\_TCD$, $S4\_COQ$, $S5\_GNQ$, $S6\_GAB$\}, where $S1\_CMR \rightarrow Cameroon$, $S2\_CAF \rightarrow CentralAfricaRepublic$, $S3\_TCD \rightarrow Chad$, $S4\_COQ \rightarrow Congo$, $S5\_GNQ \rightarrow EuatorialGuinea$, and $S6\_GAB \rightarrow Gabon$ are the quality metric function that will evaluate and measure a DQM f_k for each value (system) of an attribute e_i in the data sample s_i and returns 1 if correct, otherwise, 0. Each s_i function will determine whether the attribute's value complies with the f_k constraints. For instance, the range of values between 0 and 100 is what is meant when defining the metric accuracy of an attribute; otherwise, it is inaccurate. Similarly, the system can refer back to the data source to satisfy the data constraint. After this process, all the observed datasets from various countries will be authenticated by Benford's law to define whether the dataset is accurate before storing in the centralized CDM of the region. The metric f_k will be evaluated t measure if all the attributes individually are f_k satisfied by the condition. This is done for each instance of the data sample s_i as represented in Fig. 4 followed by Algorithms 1 and 2.

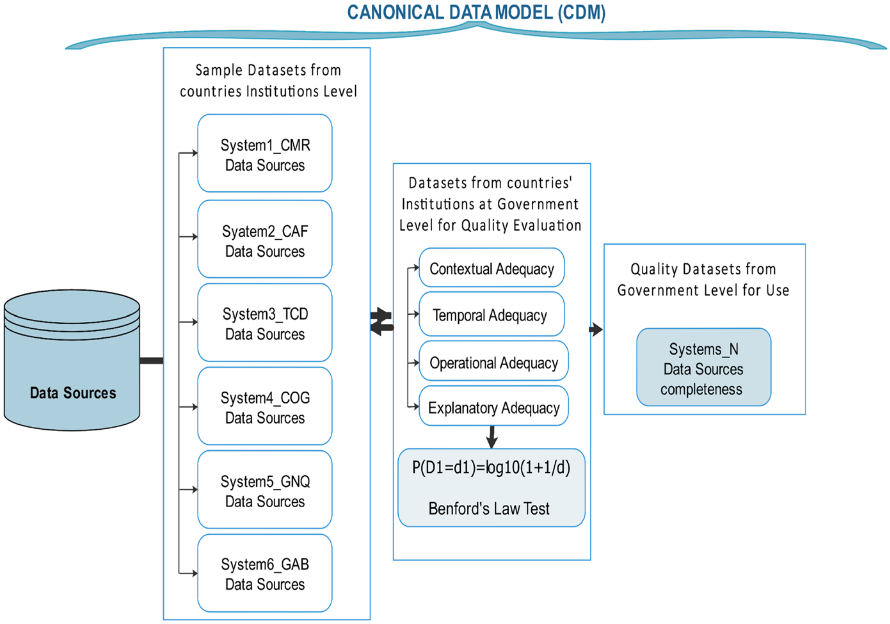

Fig. 4 Data quality model assessment

```
Algorithm 1 Data quality model measurement
Input: Data Source $\mathbb{D}$={$S1\_CMR$,
$S2\_CAF$,$S3\_TCD$,$S4\_COQ$,$S5\_GNQ$,$S6\_GAB$}
Output: Result : Cleaned Target Data Source $\mathbb{D}$
Let $ds$ an original Dataset with size $S$ and $N$ Observation($N \sim S$);
Let $s(c(S))$ the sample size with $s$¡$S$;
Let $\mathbb{D}$ a set of DQMA, $\mathbb{D}$={$S1\_CMR$,
$S2\_CAF$,$S3\_TCD$,$S4\_COQ$,$S5\_GNQ$,$S6\_GAB$};
Let $n$ samples $s_i$ with size $s$ and $M$ Observation (M~s);
Let $f_k$ a metric functions(completeness, accurracy, consistency);
Let acc ← 0 counter correct valid attribute (when$f_k$ is true acc=acc+1);
Let $\mathbb{D}' = \{\mathbb{D}'S_0, \ldots, \mathbb{D}'S_i, \ldots, \mathbb{D}'S_n\}$ be the cleaned datasets
for each iteration i in [0 - n] do
    Produce sample $S_i$ with size $S$ from $S_k$
    for each iteration i in [0 - n'] do
        // Produce a sample $s_{ij}$ of size $S$ from sample $\mathbb{D}'i$
        for each DQMA metric function tuple ($d_k$,$f_k$) do
            for each attribute ($e_{ij}$) do
                for each eij(x) s values do
                    If $(f_k(e_{ij}(x), value) == 1)$ acc←acc+1
                end
                Calculate the scores vector DQMA($f_k, d_s, e_{ij}, \mathbb{D}'S$)=acc/N acc←0 //
                 counter of correct valid attribute value($s_k, f_k$)
            end
            // DQMA $ds$ computed ∀ attributes for a sample $ds_{ij}$
        end
        // DQMA $\mathbb{D}'ijk$ is the $s_k$ scores for attributes $e_{ij}$ for a sample $\mathbb{D}'S_{ij}$;
         $Q_{ijk}$ sum of all $sk$ scores for attributes $e_{ij}$ for $\mathbb{D}'S_{ij}$
    end
    $Q_{ik} += 1/n'(Q_{ijk})$
end
// $Q_k$ is the mean of all $Q_{ik}$ for a specific $sk$
 $Q_k += 1/n(Q_{ik})$
```

4.2.2 Mathematical model of COVID-19

A deterministic compartmental model is used to simulate the COVID-19 pandemic disease propagation inside a specific contaminated territory C [32–34]). **Scenario** The corona is a mathematical consequence of severe acute respiratory syndrome occurring in a contaminated area or unit. The contiguous Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) in humans and animals account for H^∞ instances of contagious infection on the unit or territory C. Because each infected person has a maximum ideal made up of functions f with $f(z) = 0$, it contains an open subspace C in its spreading S (maximal ideals). Since the spreading is a compact space and the subspace C is not, the subspace C cannot encompass the complete spreading S.

Formulation: If $f_1, \dots, f_n \in H^\infty$, the bounded set of infection cases on the territory C in a complex plane. And if $|f_1(z)| + \cdots + |f_n(z)| \geq \delta > 0$ in C, then there $\exists g_1 \dots g_n \in H^\infty$ with $\sum f_j g_j = 1$ in C. H^∞ is a vector space under pointwise SARS and the MERS in animals and humans with the norm $\|f\| = sup|f(z)|, z \in C$. Since the pointwise product of two bounded epidemic diseases is bounded, there is a dual space H^∞ in the SARS space, which consists of all continuous linear mappings of infection cases in divers countries C from H^∞. Functions $\lambda \in H^\infty$ have the added property of being multiplicative, $\lambda(fg) = \lambda(f)\lambda(g)$. We designate the set of these multiplicative linear functionals by M. For instance, if $\alpha \in D$, then $\lambda(f) = f(\alpha)$ is in M is called a point evaluation. The set H^∞ consists of topology $\lambda\alpha$ that converges to λ if each $f \in H^\infty$, and the numbers $\lambda\alpha(f)$ converge to $\lambda(f)$. The Corona theorem stipulates "point evaluations" are dense in this topology, and Carleson [35] demonstrated that M inherits the mechanism of convergence from H^∞ by Kakutani in 1941 [36].

4.3 4A levels data quality model

Solutions for big data might be thought of as complete information systems. In this sense, our big data input targets are concerned with the transactional and analytic Covid-19 data quality in the CEMAC region countries. Our objectives are not focused on the consequences of big data computing, which quality must be assessed using divers models. An important idea of the model and its relationships is shown in Fig. 5. This input data quality level is evaluated within the data quality standard, such as ISO/IEC 25012 [22]. The proposed data quality model relies on the data quality model represented in the standard ISO/IEC 25010 [22, 28]. Data understanding is a product to which data can be applied in accordance with the quality interpretation established by ISO/IEC 25010. The four adequacy levels (4A) data quality model presented in this study is intended to close the hole and evaluate the COVID-19 data quality model.

This new model is proposed to show how sound and relevant the data is from a quality standpoint for the intended data assessment objective. We suggest that the main key point to evaluate the degree of data quality concern on COVID-19 data is the adequacy of data quality to be analyzed. The ability to be adequate for a need, purpose, or demand is called adequate. The ability of data to be sufficient to achieve the objectives and aims of the analysis is known as data adequacy. The classification of DM Strong et al. [37] served as the starting point for identifying the key attributes of our data quality model. To meet the characteristics of Representational, Accessibility, and Intrinsic, these categories were split into two features for the big dataset: Contextual Adequacy and Operational Adequacy. This simplification was made mostly due to the requirement that data be processable using the tools and technology available for big data analysis. These three criteria fall under the definition of a single quality that we refer to as Operational Adequacy. Regarding Contextual Adequacy, we concede that the temporal elements of the context are adequate in terms of context. However, a solitary evaluation of the temporal elements was deemed necessary due to the real-time analysis'

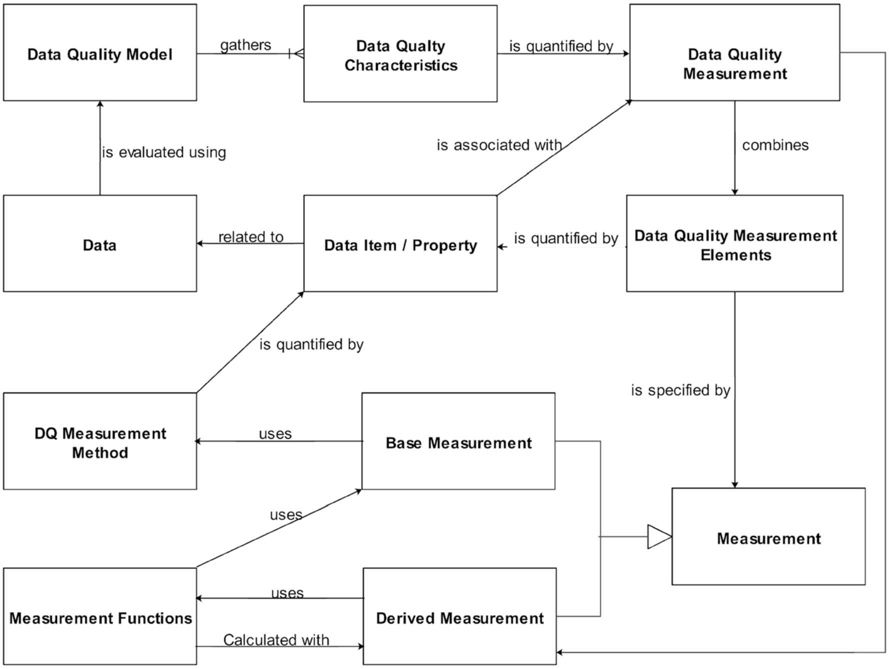

Fig. 5 The main concept of the data quality model assessment [22]

expanding significance. The Contextual category was consequently divided into Temporal Adequacy and Contextual Adequacy. As a result, we pinpoint the following four crucial data quality traits as being crucial in big data analysis: Contextual, Temporal, Operational, and Explanatory Adequacy. Subsequently, the definition of each trait from the 4A levels data quality model is then given:

Regardless of type (e.g., structured vs. unstructured), quantity, or inflow velocity, **Contextual Adequacy** is the capacity of datasets to be employed within the same domain of interest for the study. Therefore, information must be:

- Relevant and comprehensive: data must be sufficient and appropriate for the task at hand (such as in the case of a big data analysis);
- Data must be understood in accordance with the context presented and be free of inconsistencies caused by duplications to be unique and semantically compatible;
- In big data analysis's framework, data characterize real things in a semantic precise manner;
- Credible: Analysts should assess the context-specific degrees of data integrity (the trustworthy of data source);
- Confidential: The data must be accessible to the same group of people who are allowed to do the analysis;
- Comply with the stated rules and regulations.

Data that meets the **Temporal Adequacy** criteria lies within an appropriate time frame for the analysis, such as contemporaneous, similar in age, or span a certain period for historical data. It is significant to note that this definition only addresses the temporal characteristics of the data themselves and does not address the temporal aspects of data processing. Since there are numerous ways to interpret this, the data processing must be:

- time-concurrent: refers to events that took place at appropriate times or similar (for instance, if a study is centered on a past occurrence, data must link to related and contemporaneous events);
- current: The age of data should be consistent. Combining data with varied degrees of accuracy may not always produce solid analysis;
- timely updated: data must be correctly updated for the current task in order to be used in the analysis;
- frequent: The data used to generate insights from any trend analysis are typically tied to future time slots for records (necessary frequencies).
- time-consistent: Data must be coherent with the represented time and free of any incoherence to be considered (e.g., disordered events, impossible dates, etc.).

The degree of data processing for intended analysis by a suitable technologies without omitting a piece of data from the analysis is called **Operational Adequacy**. This demonstrates that adequate and suitable resources are available for the analysis (for instance, similar data types, equivalently expressed data attributes, etc.). The performance and cost-effectiveness of the 4Vs must both be considered. Therefore, information in the multiple datasets should:

- be easily recovered, available for analysis, and accessible;
- be granted permission for the intended uses;
- be expressed using same data types, with an equivalent degree of precision, and be portable;
- have a convincing representation to save money;
- provide an audit trail that permits tracking changes and accesses;

Explanatory Adequacy states that data quality knowledge must abide by a set of global standards for adequate quality data. This indicates that there is a major and appropriate policy for the use of the available resources. It is important to take into account the 4Vs (volume, velocity, variety, and veracity) in terms of cost-effectiveness and performance.

As a result, the various datasets should adhere to the following operational issues:

- Data quality is still a problem. As more data is collected, it becomes harder to analyze everything accurately and reliably;
- It's crucial to have accurate data descriptions (metadata). It's important to organize and process data;
- Interpretation of data remains mostly of an art than a science;
- The data perspective representation is challenging or difficult to master;

- The analysis of the data in real-time is somewhat out of date;
- It is preferable to have a long-term data custody policy, and this policy must be followed.

Table 2 shows the data quality features that may influence the four adequacy levels as data quality model.

4.4 4A quality in canonical data model(CDM)

A sort of data model known as a canonical data model (CDM) displays data items and associations in the most straightforward manner. It is sometimes referred to as a common data model since it enables data exchange between systems of any technology. A CDM is widely used in system or database integration operations when data are shared across several platforms and operating systems. In terms of adequacy levels, it essentially gives an organization the power to develop and disseminate a single definition of its complete data unit. Identification of all entities, their characteristics, and the connections between them is necessary for the design of a CDM. When data units are shared between various information system platforms during integration processes, a CDM's reputation is most readily apparent. It presents or defines data using a complete data format, enabling data sharing between various applications easier. These six entities in Fig. 6 represent the six CEMAC region countries. The acronyms for the entities are defined in Table 3 below.

Table 2 Data quality model evaluation for big dataset based on 4A levels

Data quality features	Contextual Adequacy	Temporal Adequacy	Operational Adequacy	Explanatory Adequacy
Accuracy	x			
Consistency	x	x		x
Credibility	x			x
Currentness		x		
Completeness	x			
Accessibility			x	
Efficiency			x	x
Confidentiality	x		x	x
Precision			x	
Compliance	x			x
Traceability			x	x
Availability			x	x
Understandability	x			x
Portability			x	x
Recoverability			x	x

Fig. 6 Canonical data model

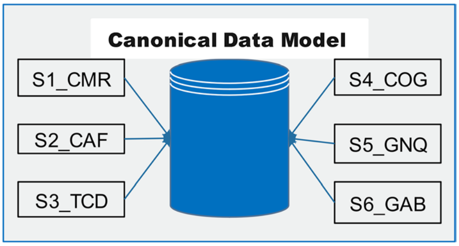

Table 3 Analytical-based systems for CEMAC region countries

Attributes	Description
S1_CMR	System 1 for the Republic of Cameroon
S2_CAF	System 2 for the Central Africa Republic
S3_TCD	System 3 for the Republic of Chad
S4_COG	System 4 for the Republic of Congo
S5_GNQ	System 5 for the Republic of Equatorial Guinea
S6_GAB	System 6 for the Republic of Gabon

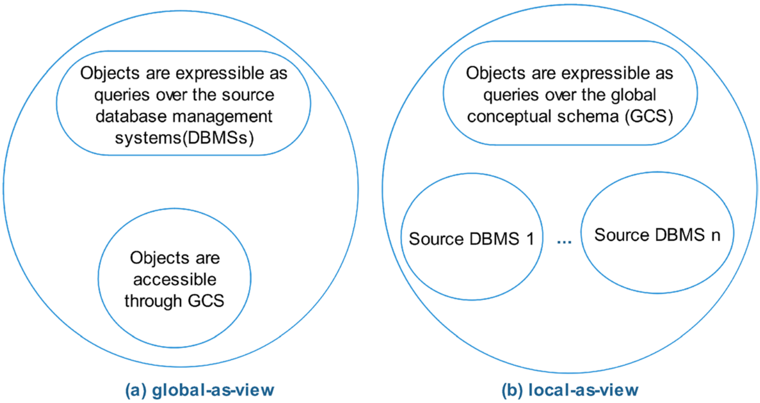
Objects are expressible as queries over the source database management systems(DBMSs)
Objects are accessible through GCS
(a) global-as-view
Objects are expressible as queries over the global conceptual schema (GCS)
Source DBMS 1
...
Source DBMS n
(b) local-as-view

Fig. 7 GAV and LAV mapping

4.4.1 Canonical data model design methodology

The CDM design comprises of the procedure by which information from participating databases from these six CEMAC region countries can be defined up-front. In this context, local-as-view (LAV) and global-as-view (GAV) factors can be used to categorize the association between the global conceptual or mediated schema (GCS) and the local conceptual schemas (LCSs) [38]. Because the GCS is provided in LAV systems, each LCS is viewed as a view definition over the GCS. In the GAV systems, the GCS is described as a group of views over the LCSs. These points of view classify the various ways that the LCS elements can be converted into the GCS elements. Their distinctions can be demonstrated, for example, by comparing the outcomes of each system [39]. The set of objects described in the GCS is the only set that may be accessed by a query in GAV, even though local database management systems (DBMSs) may have a far greater range of items (Fig. 7a).

In contrast, the GCS definition may be more comprehensive. The objects in the local database management systems (DBMSs) in LAV limit the results (Fig. 7b). So it might be required to cope with incomplete answers in LAV systems. A hybrid of these two strategies known as global–local-as-view (GLAV) [40] has also been put forth, in which the relationship between GCS and LCSs is defined by the employment of both LAV and GAV.

4.4.2 Canonical mapping generation and illustration

The process of creating a mapping starts with the source LCS, the destination GCS, a set of M schema matches, and a set of queries that, when run, produce GCS data instances from the data source.

Let us consider for more concrete illustration by referral to the canonical relational representation that we have adopted. A set of relations $S = S_1 \cdots + S_m$ make up the source LCS under examination. A set of global or target relations, $T = T_1 \ldots T_n$, and M consists of a set of schema match rules make up the GCS. In order to produce data for each T_k from the source relations, we are creating a query Q_k for each T_k that is described on a subset of relations in S. This is accomplished iteratively via an algorithm used in [41] that takes each T_k into account in turn. It begins with $M_k \subseteq M$, where M_k is the set of rules that only apply to the characteristics of T_k, and subdivides it into $\{M_k^1, \ldots, M_k^s\}$, where each subset defines a potential method for computing the values of T_k. For example, we will use an illustrative below to demonstrate the algorithm. Since we have six (6) CEMAC region countries represented as follows:

System 1 → Republic of Cameroon ($S1\_CMR$);
System 2 → Central Africa Republic($S2\_CAF$);
System 3 → Republic of Chad ($S3\_TCD$);
System 4 → Republic of Congo ($S4\_COQ$);
System 5 → Republic of Equatorial Guinea ($S5\_GNQ$);
System 6 → Republic of Gabon ($S6\_GAB$).

Source relations (LCS):

$S1\_CMR(A_1, A_2)$;
$S2\_CAF(B_1, B_2, B_3)$;
$S3\_TCD(C_1, C_2, C_3)$;
$S4\_COQ(D_1, D_2, D_3)$;
$S5\_GNQ(E_1, E_2)$;
$S6\_GAB(F_1, F_2)$.

Target relation (GCS): $T(W_1 W_2 W_3 W_4 W_5 W_6)$

4.5 Benford's law

Benford's law is an observation of finding the first digits of the numbers in real-world data sets. Intuitively, one may anticipate that the distribution number of these leading digits would be uniformly distributed, giving any digits from 1 to 9 an equal chance of showing up. It frequently happens that 1 occurs more frequently than 2, 2 more frequently than 3, and so on [42]. Benford's law is condensed in this comment. To be more precise, the Law estimates the frequency of leading digits using base-10 logarithms that predict specific frequencies, which decrease as the digits increase from 1 to 9. Benford's law specifies a statistical distribution for many datasets' first and higher-order digits. Numbers are anticipated to naturally follow the predicted digits pattern under very general circumstances. Conversely, any deviation from the Benford distribution could point to an external alteration of the anticipated pattern brought on by data fraud or manipulation. Many statistical tests are available to evaluate the Benford conformity of a sample. However, in some real-world scenarios, there can be concerns about the dependability of the data due to the small amount available for analysis. Publications in science, technology, and business are currently being checked twice. It's an unfortunate situation, but true development. We are surrounded by statistical models and conclusions in this age of big data. These studies and models significantly impact our society and their findings in various areas, including healthcare, economics, social interaction, and technology research. The fact that basic science research publications, which ought to report the pure, objective truth, are not exempt from such dishonesty and fraud is all the more depressing. We require strong analytical methods and efficient screening procedures to assess these datasets' validity. As a result, a real-world example is given to show how useful and effective a sounding testing Benford compliance test for these various sample datasets is for anti-fraud investigations.

The rule of the first digits, also known as Benford's law [42], predicts that for a given collection of integers, those with a first digit of "1" will occur more frequently (30.103%) than those with a first digit of "2" to "9" (17.609%, 12.494%, 9.691%, 7.918%, 6.695%, 5.799%, 5.115%, and 4.576%). The fact that in a variety of situations, the frequency with which objects form "naturally" is an inverse function of their size can be used to explain why Benford's law fits empirical evidence so well. More often than little objects, which more frequently than large ones, are small objects [5]. This law is formulated as follow:

suppose that $\langle x \rangle = x - \lfloor x \rfloor$ for $x \in \mathbb{R}$, where $\lfloor x \rfloor$ denotes the "floor fiction", i.e $\lfloor x \rfloor = \max\{n \in \mathbb{Z} : n \leq x\}$. It is ostensible that $\langle x \rangle$ stands for fractional part of x and it is also designate with {x} in [43]. More so, each not-null $x \in \mathbb{R}$ the so-called "significant function" $S :\rightarrow \mathbb{R}[1, 10[$ is provided by $S(x) = 10^{\langle log_{10}|x| \rangle}$, while $S(0) := 0$ & $x = 0$ [43]. For suitability, $s(x) = log_{10}S(x)$ and the first significant digit x of $D_1(x)$, may be restructured in terms of significant function (i.e. $D_1(x) = \lfloor S(x) \rfloor$, and kth significant digit x of $D_1(x)$, may be generally reconstruct as $D_k(x) = \lfloor 10^{k-1}S(x)\rfloor - 10\lfloor 10^{k-2}S(x)\rfloor$ for $k = 2, 3, \dots$ [43]). Since S(x)=$\sum_{l\in\mathbb{N}} 10^{-l+1}D_1(x)$, it holds $\{x \in \mathbb{R} : D_l(x) = d_k\} = x \in \mathbb{R} : S(x) \in [10^{l-k}C_{d_l,\dots,d_k}, 10^{l-k}(C_{d_l,\dots,d_k} + 1)[$, where $C_{d_l,\dots,d_k} = \sum_l^k 10^{k-l}d_l$, while $d_1 \in 1,\dots,9$ & $d_l \in 1,\dots,9$ for $l = 2,\dots,k$.

Let a random variable X be the probability space (Ω,Γ,P). The joint probability distribution based on the prior remarks of the random vector $(D_l(X),\dots,D_k(X))$ by $P(D_l(X) = d_l,\dots,D_k(x) = d_k\} = PS(x) \in [10^{l-k}C_{d_l,\dots,d_k}, 10^{l-k}(C_{d_l,\dots,d_k} + 1)[$, where $C_{d_l,\dots,d_k} = \sum_l^k 10^{k-l}d_l$, while $d_1 \in 1,\dots,9$ and $d_l \in 1,\dots,9$ for $l = 2,\dots,k$. In particular, it shows that $P(D_l(X) = d_l) = P(S(X) \in [d_l, d_l + 1[)$. According to [43, 44], the random variable is said to be Benford if $P(S(X) \leq t) = \log_{10} t$ for $t \in$[1,10[and it obviously equivalent to $P(S(X) \leq u) = u$ for u∈[0,1[. Thus, the joint probability distribution of the random vector $(D_l(X),\dots,D_k(X))$ is given by $P(D_l(X) = d_l,\dots,D_k(x) = d_k) = \log_{10}(\frac{C_{d_l,\dots,d_k}+1}{C_{d_l,\dots,d_k}})$. Moreover, for $k = 1$, this expression reduces to $P(D_l(X) = d_l) = \log_{10}(\frac{d_1+1}{d_1})$, which is the famous result originated by Mewcomb in 1881 [44] and consequently rediscovered by Benford in 1938 [42]. On the basis of [44], the random variable X is assumed to be the generalized Benford's law (GBL) with the parameter α if $P(S(X) \leq t) = log_{10}t \;\; \alpha = 0$ and $\frac{t^{\alpha}-1}{10^{\alpha}-1} \;\; \alpha \neq 0$. For $t \in [1, 10[$ or equivalently $P(S(X) \leq u) = u, \;\; \alpha = 0$ and $\frac{t^{\alpha u}-1}{10^{\alpha}-1} \;\; \alpha \neq 0$ for $u \in [0, 1[$.

```
Algorithm 2 Benford's Law Algorithm
Input: $\mathbb{D}$ = A big dataset with Variables $(x_1, x_2, ..., x_n)$
Output: $P(D_l(X) = d_l) = log_{10}(\frac{d_1+1}{d_1})$
Let $S[i] : \mathbb{R} \rightarrow [1, 10[$ by $S[i] = 10^{\lfloor log_{10}|x| \rfloor}$ while S(0):=0 & x=0
  Let r.v defined the probability space (Ω,Γ,P)
  joint pdf of random vector $(D_l(X), ..., D_k(X))$ by $P(D_l(X) = d_l, ..., D_k(x) = d_k\} = PS(x) \in [10^{l-k}C_{d_l,...,d_k}, 10^{l-k}(C_{d_l,...,d_k} + 1)[$,
  where $C_{d_l,...,d_k} = \sum_l^k 10^{k-l}d_l,'$
  while $d_1 \in 1, ..., 9$ & $d_l \in 1, ..., 9$ for $l = 2, ..., k$.
  Print $P(D_l(X) = d_l) = log_{10}(\frac{d_1+1}{d_1})$
  end
```

5 Experimental analysis

In this section, we demonstrate our experiments to evaluate the data quality of the COVID-19 big dataset.

5.1 Experiments setup

The experiments were conducted on a Computer Legion System (GPU) with a Processor Intel(R)Core (TM) i7-8750H CPU @ 2.2GHz, 2208 MHz, 6 cores, 12

Logical Processor(s), 16GB of memory, 512GB SSD, 1TB HDD, NVIDIA GTX 1060 and Windows 10 Operating System. The approach was put into practice using the Python programming language, Anaconda version 3, and Microsoft Excel [45] (Table 4).

5.2 Datasets

To gather evidence regarding the level of the epidemiological surveillance system's performance, we used Our World in Data by Hannah Ritchie et al. [46], and datasets for the six CEMAC region countries from March 6, 2020, to June 22, 2022, only were extracted. A sample consisting of 4974 instances, 66 observations, 161,694 missing observations, and 0 data duplication. The datasets were updated daily from the WHO situation reports [46]. In this situation, reports were the number of confirmed cases, suspected cases, and death cases as labels in the past 24 h, and cumulated confirmed cases and death cases in each country, region, and city. The log-likelihood ratio test was used to assess cumulative cases gathered from various nations and reported by the WHO in terms of how well they adhered to the data quality model and the distribution of Benford's law. More details about the dataset are available at: https://ourworldindata.org/ or https://covid19.who.int/WHO-COVID-19-global-data.csv (Accessed on 15th June 2022).

5.3 Evaluation of data quality levels

Data Quality Model Assessment (DQMA) suitability is obtained from the characteristics of the data profiling function. The profiling data function extracted the results from various countries' data collection. It sets at the scale of input and output values, i.e., input values [0–100] and output values [1–5]. The output profiling is estimated in ranges. The range is a vector that indicates the maximum value or percentage of items in each level to get the output value. For example, suppose one evaluates a

Table 4 Notation definition

Notation	Definition
acc	Accuracy
r.v	Random variable
DQMA	Data quality model assessment
DQ	Data Quality
S1_CMR	System number 1 for the Republic of Cameroon
S2_CAF	System number 2 for the Central Africa Republic
S3_TCD	System number 3 for the Republic of Chad
S4_COG	System number 4 for the Republic of Congo
S5_GNQ	System number 5 for the Republic of Equatorial Guinea
S6_GAB	System number 6 for the Republic of Gabon
∃	There exists

system and gets the following values of the characteristics: 23.98, 33.37, and 67.51. The value suitability from the characteristics is classified into levels depending on their value, as shown in Table 5. Therefore, the value 23.98 belongs to level 1, 33.37 belongs to level 2, and 67.51 belongs to level 3. It is also shown that the quality of these datasets reaches the required level (see Table 5). Table 5 presents the profiling

Table 5 Data quality model adequacy levels of the six CEMAC region countries datasets

Accuracy (%)	DQ levels	Quality values	DQMA (%)
Republic of Cameroon			
	1	0–24	
	2	25–49	
65.77	3	50–74	97
	4	75–94	
	5	95–100	
Central Africa Republic			
	1	0–24	
	2	25–49	
66.32	3	50–74	97
	4	75–94	
	5	95–100	
Republic of Chad			
	1	0–24	
	2	25–49	
66.62	3	50–74	97
	4	75–94	
	5	95–100	
Republic of Congo			
	1	0–24	
	2	25–49	
68.02	3	50–74	97
	4	75–94	
	5	95–100	
Republic of Equatorial Guinea			
	1	0–24	
	2	25–49	
67.51	3	50–74	97
	4	75–94	
	5	95–100	
Republic of Gabon			
	1	0–24	
	2	25–49	
66.63	3	50–74	97
	4	75–94	
	5	95–100	

assessment levels performance indicated by our model and the accuracy of the data quality of each country from the six CEMAC region countries.

5.4 Discussions

It is practically challenging to produce high-quality statistics in this pandemic era without extensive cooperation/collaboration between health authorities, healthcare providers, and researchers from other sectors. Due to the COVID-19 pandemic's urgency, datasets of poorer quality may be used, endangering the validity of the conclusions and leading to biased evidence. Poor decision-making or not using data to inform decisions could be the results. Access to high quality health data is one of the methodological difficulties connected with evaluating COVID-19 data during the pandemic, and various data quality issues have been discussed. However, to our knowledge, no existing study systematically evaluates the data quality problems in the datasets provided by the national surveillance systems of the six CEMAC region countries for research purposes during the COVID-19 pandemic. Despite this being a global issue, this study solely uses WHO datasets for the six CEMAC region countries.

The existing data quality evaluation methods for Covid-19 big datasets focus only on outbreak prediction and data quality evaluation in the COVID-19 surveillance systems data. Researchers also made an analysis of national COVID-19 reporting methods in 54 African nations by report type, frequency, data content, and reporting systems were compared. The patient demographics and morbidities, the capacity of the healthcare system, and diagnostic testing were all compared as reporting metrics. Table 5 presents the assessment levels performance indicated by our model and the accuracy of the data quality of each country from the six CEMAC region countries. All these accuracies (65.77%, 66.32%, 66.62%, 68.02%, 67.51% and 66.63%) are at data quality level 3 according to the quality value range [50–74], which is fairly acceptable for this situation. Table 6 shows the sample extracted cases distribution data of the six CEMAC region countries with its issues and the starting time of COVID-19 in these countries. Table 7 presents the performance of Benford's law on each country dataset. As shown, the distribution data is not fitted properly with respect to Benford's law. We conclude that the six CEMAC region countries' datasets lack quality. Our results indicate a significant disparity in the quality of COVID-19 data reporting across those six CEMAC region countries. The records range from [0–100] for the Quality value, [1–5] for the Data Quality levels, and Benford's law for an objective and fast way to access the performance of the surveillance systems and data collected during the epidemics. The result in Table 5 varies from 65.77 to 68.02% for the six CEMAC region countries, which lies at data quality Level 3 (fair). In addition, we identify the non-correlation of the data with Benford's law, as shown in Table 7 and Fig. 8. The blue bars represent Benford's law results obtained from each country dataset compared with the orange bars, which is the standard Benford's law distributed leading digits lying between 1 and 9.

It is critical everywhere that scientific data be produced to aid in the management of the COVID-19 epidemic. However, if the quality of the datasets is poor, the

Table 6 Extracted sample of CEMAC region countries datasets from the WHO from March 6, 2020, to June 22, 2022

Row	Date	Code	Area	Country	Total cases	New cases	Deaths	New deaths	Samples	Missing	Number of variables	Population
30,261	06-03-2020	CMR	Arifica	Cameroon	1	1	–	–	836	28,649	68	27,224,262
30,262	07-03-2020	CMR	Africa	Cameroon	1	0	–	–	836	28,649	68	27,224,262
30,263	08-03-2020	CMR	Africa	Cameroon	2	1	–	–	836	28,649	68	27,224,262
30,264	09-03-2020	CMR	Africa	Cameroon	2	0	–	–	836	28,649	68	27,224,262
30,265	08-03-2020	CMR	Africa	Cameroon	2	0	–	–	836	28,649	68	27,224,262
33,631	15-03-2020	CAF	Africa	Central Africa Republic	1	1	–	–	827	27,889	66	4,919,987
33,632	16-03-2020	CAF	Africa	Central Africa Republic	1	0	–	–	827	27,889	66	4,919,987
33,633	17-03-2020	CAF	Africa	Central Africa Republic	1	0	–	–	827	27,889	66	4,919,987
33,634	18-03-2020	CAF	Africa	Central Africa Republic	1	0	–	–	827	27,889	66	4,919,987
33,635	19-03-2020	CAF	Africa	Central Africa Republic	1	0	–	–	827	27,889	66	4,919,987
34,459	19-03-2020	TCD	Africa	Republic of Chad	1	1	–	–	824	27,502	66	16,914,985
34,460	20-03-2020	TCD	Africa	Republic of Chad	1	0	–	–	824	27,502	66	16,914,985
34,461	21-03-2020	TCD	Africa	Republic of Chad	1	0	–	–	824	27,502	66	16,914,985
34,462	22-03-2020	TCD	Africa	Republic of Chad	1	0	–	–	824	27,502	66	16,914,985
34,463	23-03-2020	TCD	Africa	Republic of Chad	1	0	–	–	824	27,502	66	16,914,985
34,464	24-03-2020	TCD	Africa	Republic of Chad	3	2	–	–	824	27,502	66	16,914,985
34,465	25-03-2020	TCD	Africa	Republic of Chad	3	0	–	–	824	27,502	66	16,914,985
34,466	26-03-2020	TCD	Africa	Republic of Chad	3	0	–	–	824	27,502	66	16,914,985
34,467	27-03-2020	TCD	Africa	Republic of Chad	3	0	–	–	824	27,502	66	16,914,985
34,468	28-03-2020	TCD	Africa	Republic of Chad	3	0	–	–	824	27,502	66	16,914,985
34,469	29-03-2020	TCD	Africa	Republic of Chad	3	0	–	–	824	27,502	66	16,914,985
34,470	30-03-2020	TCD	Africa	Republic of Chad	5	2	–	–	824	27,502	66	16,914,985
34,471	31-03-2020	TCD	Africa	Republic of Chad	7	2	–	–	824	27,502	66	16,914,985
34,472	01-04-2020	TCD	Africa	Republic of Chad	7	0	–	–	824	27,502	66	16,914,985

Table 6 (continued)

Row	Date	Code	Area	Country	Total cases	New cases	Deaths	New deaths	Samples	Missing	Number of variables	Population
34,473	02-04-2020	TCD	Africa	Republic of Chad	8	1	–	–	824	27,502	66	16,914,985
34,474	03-04-2020	TCD	Africa	Republic of Chad	8	0	–	–	824	27,502	66	16,914,985
34,475	04-04-2020	TCD	Africa	Republic of Chad	9	1	–	–	824	27,502	66	16,914,985
34,476	04-04-2020	TCD	Africa	Republic of Chad	9	1	–	–	824	27,502	66	16,914,985
38,633	15-03-2020	COQ	Africa	Republic of Congo	1	1	–	–	824	26,451	66	5,657,017
38,634	16-03-2020	COQ	Africa	Republic of Congo	1	0	–	–	824	26,451	66	5,657,017
38,635	17-03-2020	COQ	Africa	Republic of Congo	1	0	–	–	824	26,451	66	5,657,017
38,636	18-04-2020	COQ	Africa	Republic of Congo	1	0	–	–	824	26,451	66	5,657,017
38,637	19-04-2020	COQ	Africa	Republic of Congo	3	2	–	–	824	26,451	66	5,657,017
38,638	20-03-2020	COQ	Africa	Republic of Congo	3	0	–	–	824	26,451	66	5,657,017
38,639	21-03-2020	COQ	Africa	Republic of Congo	3	0	–	–	824	26,451	66	5,657,017
38,640	22-03-2020	COQ	Africa	Republic of Congo	3	0	–	–	824	26,451	66	5,657,017
38,641	23-03-2020	COQ	Africa	Republic of Congo	4	1	–	–	824	26,451	66	5,657,017
38,642	24-03-2020	COQ	Africa	Republic of Congo	4	0	–	–	824	26,451	66	5,657,017
38,643	25-03-2020	COQ	Africa	Republic of Congo	4	0	–	–	824	26,451	66	5,657,017
38,644	26-03-2020	COQ	Africa	Republic of Congo	4	0	–	–	824	26,451	66	5,657,017
38,645	27-03-2020	COQ	Africa	Republic of Congo	4	0	–	–	824	26,451	66	5,657,017
38,646	28-03-2020	COQ	Africa	Republic of Congo	4	0	–	–	824	26,451	66	5,657,017
38,647	29-03-2020	COQ	Africa	Republic of Congo	19	15	–	–	824	26,451	66	5,657,017
38,648	30-03-2020	COQ	Africa	Republic of Congo	19	0	–	–	824	26,451	66	5,657,017
38,649	31-03-2020	COQ	Africa	Republic of Congo	19	0	–	–	824	26,451	66	5,657,017
38,650	01-04-2020	COQ	Africa	Republic of Congo	19	0	–	–	824	26,451	66	5,657,017
52,425	15-03-2020	GNQ	Africa	Equatorial Guinea	1	1	–	–	827	26,868	66	1,449,891
52,426	16-03-2020	GNQ	Africa	Equatorial Guinea	1	0	–	–	827	26,868	66	1,449,891

Table 6 (continued)

Row	Date	Code	Area	Country	Total cases	New cases	Deaths	New deaths	Samples	Missing	Number of variables	Population
52,427	17-03-2020	GNQ	Africa	Equatorial Guinea	1	0	–	–	827	26,868	66	1,449,891
52,428	18-03-2020	GNQ	Africa	Equatorial Guinea	4	3	–	–	827	26,868	66	1,449,891
52,429	19-03-2020	GNQ	Africa	Equatorial Guinea	6	2	–	–	827	26,868	66	1,449,891
63,466	14-03-2020	GAB	Africa	Gabon	1	1	–	–	829	27,667	68	2,388,992
63,467	15-03-2020	GAB	Africa	Gabon	1	0	–	–	829	27,667	68	2,388,992
63,468	16-03-2020	GAB	Africa	Gabon	1	0	–	–	829	27,667	68	2,388,992
63,469	17-03-2020	GAB	Africa	Gabon	1	0	–	–	829	27,667	68	2,388,992
63,469	18-03-2020	GAB	Africa	Gabon	1	0	–	–	829	27,667	68	2,388,992

Table 7 Benford's law evaluation for the six CEMAC region countries

Digit	Frequency	Percentage (%)	Bendford's law (%)	Benford's law expected	$(O-E)^2/E$	Cumulative (%)	Benford's cumulative (%)	Chi-square	p value
Republic of Cameroon									
1	349	41.75	30.10	2777.131887	2122.990445	41.75	30.10		
2	169	20.22	17.61	4747.538319	4415.554279	61.96	47.71		
3	44	5.26	12.49	6691.279444	6603.568776	67.22	60.21		
4	28	3.35	9.69	8626.559569	8570.650451	70.57	69.90		
5	16	1.91	7.92	10,558.05562	10,526.07987	72.49	77.82	93,296.88996	0
6	29	3.47	6.69	12,487.52934	12,429.59669	75.96	84.51		
7	50	5.98	5.80	14,415.79467	14,315.96809	81.94	90.31		
8	123	14.71	5.12	16,343.28006	16,098.20576	96.65	95.42		
9	28	3.35	4.58	18,270.23269	18,214.2756	100.00	100.00		
Total	836	100.00	100.00	94,917.4016	93,296.88996	–	–		
Central Africa Republic									
1	353	42.68	30.10	2747.234534	2086.592511	42.68	30.10		
2	11	1.33	17.61	4696.428457	4674.454221	44.01	47.71		
3	32	3.87	12.49	6619.244139	6555.398839	47.88	60.21		
4	231	27.93	9.69	8533.689908	8077.942889	75.82	69.90		
5	65	7.86	7.92	10,444.39234	10,314.79687	83.68	77.82	92,294.39875	0
6	30	3.63	6.69	12,353.09422	12,293.16708	87.30	84.51		
7	89	10.76	5.80	14,260.60071	14,083.15615	98.07	90.31		
8	13	1.57	5.12	16,167.33565	16141.34611	99.64	95.42		
9	3	0.36	4.58	18,073.54359	18067.54408	100.00	100.00		
Total	827	100.00	100.00	93,895.56355	92,294.39875	–	–		
Republic of Chad									

Table 7 (continued)

Digit	Frequency	Percentage (%)	Bendford's law (%)	Benford's law expected	$(O-E)^2/E$	Cumulative (%)	Benford's cumulative (%)	Chi-square	p value
1	138	16.75	30.10	2737.26875	6.957300045	16.75	30.10		
2	31	3.76	17.61	4679.391836	0.205368568	20.51	47.71		
3	60	7.28	12.49	6595.23237	0.545848849	27.79	60.21		
4	195	23.67	9.69	8502.733355	4.472091316	51.46	69.90	91,922.67831	0
5	123	14.93	7.92	10,406.50458	1.453802271	66.38	77.82		
6	27	3.28	6.69	12,308.28251	0.05922841	69.66	84.51		
7	159	19.30	5.80	14,208.86939	1.779240791	88.96	90.31		
8	53	6.43	5.12	16,108.68752	0.174377956	95.39	95.42		
9	38	4.61	4.58	18,007.98055	0.08018667	100.00	100.00		
Total	824	100.00	100.00	93,554.95086	15.72744488	–	–		
Republic of Congo									
1	300	36.28	30.10	2747.234534	2179.994752	36.28	30.10		
2	209	25.27	17.61	4696.428457	4287.729354	61.55	47.71		
3	51	6.17	12.49	6619.244139	6517.637084	67.71	60.21		
4	38	4.59	9.69	8533.689908	8457.85912	72.31	69.90		
5	83	10.04	7.92	10,444.39234	10,279.05193	82.35	77.82	92,285.21113	0
6	37	4.47	6.69	12,353.09422	12,279.20504	86.82	84.51		
7	41	4.96	5.80	14,260.60071	14,178.71858	91.78	90.31		
8	35	4.23	5.12	16,167.33565	16,097.41142	96.01	95.42		
9	33	3.99	4.58	18,073.54359	18,007.60384	100.00	100.00		
Total	827	100.00	100.00	93,895.56355	92,285.21113	–	–		
Republic of Equatorial Guinea									
1	335	40.51	30.10	248.9518064	29.74186742	40.51	30.10		

Table 7 (continued)

Digit	Frequency	Percentage (%)	Bendford's law (%)	Benford's law expected	$(O-E)^2/E$	Cumulative (%)	Benford's cumulative (%)	Chi-square	p value
2	13	1.57	17.61	145.6274712	120.7879666	42.08	47.71		
3	37	4.47	12.49	103.3243352	42.57387603	46.55	60.21		
4	54	6.53	9.69	80.14458076	8.528824988	53.08	69.90		
5	173	20.92	7.92	65.48289048	176.5335762	74.00	77.82	452.7288133	9.62119E−93
6	37	4.47	6.69	55.36499502	6.091810215	78.48	84.51		
7	54	6.53	5.80	47.95934015	0.7608439	85.01	90.31		
8	95	11.49	5.12	42.30313606	65.64429324	96.49	95.42		
9	29	3.51	4.58	37.84144469	2.065754754	100.00	100.00		
Total	827	100.00	100.00	827	452.7288133	–	–		
Republic of Gabon									
1	109	13.15	30.10	249.5538664	79.16282623	13.15	30.10		
2	197	23.76	17.61	145.9796538	17.83177082	36.91	47.71		
3	109	13.15	12.49	103.5742126	0.284232606	50.06	60.21		
4	188	22.68	9.69	80.33840078	144.2774543	72.74	69.90		
5	23	2.77	7.92	65.64125297	27.70020944	75.51	77.82	400.483943	1.47589E−81
6	19	2.29	6.69	55.4988886	24.00352336	77.80	84.51		
7	19	2.29	5.80	48.07532404	17.58437379	80.10	90.31		
8	90	10.86	5.12	42.40544111	53.41866461	90.95	95.42		
9	75	9.05	4.58	37.93295967	36.22088786	100.00	100.00		
Total	829	100.00	100.00	829	400.483943	–	–		

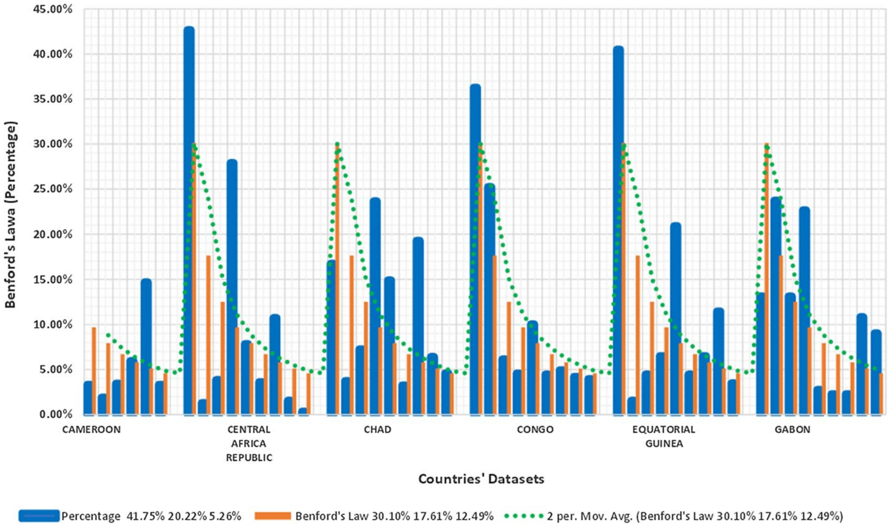

Fig. 8 Benford's law comparison with the six CEMAC region countries datasets

evidence generated can be unreliable and hence only be somewhat applicable. This issue may be especially serious when skewed conclusions are repeated throughout research using low-quality datasets provided by reputable institutions. The problem of using datasets with sub-optimal quality for research or tackling the spread of the disease during the pandemic probably occurs in many countries. Using the six CEMAC region countries' datasets, this research reports many inconsistencies and incompleteness of data (Tables 5, 6, and 7) that may overlap with scientific conclusions. Table 2 presents data quality issues detected in the used datasets at different levels of adequacies and possible solutions. The missing and inconsistent data issues might be due to the data collection mechanism and/or the database records sent to researchers or WHO. In striving to produce and provide high-quality data to the public, DQMA represents best practices for the six CEMAC region countries and other institutions, and agencies to emulate along with these proposed approaches.

6 Benefits

This study proposes a novel data quality model in the sense to help the research community and public health practitioners by allowing them to provide accurate and accessible COVID-19 data. The following are the primary advantages of this model:

1. 4A adequacy levels represent the interdependence of all sorts of attributes defined in ISO/IEC25010. As a product that meets the established requirements, quality interpretation can be used to data comprehension. The four adequacy levels (4A) data quality model is created to bridge a gap in the data quality model, empowering the assessment of the covid-19 data quality model.

2. Canonical Data Model is a design pattern for an enterprise's applications and services, which could be any app or service, to standardize on agreed data definitions or formats, reducing the number of data translations, and lowering maintenance effort and cost. Canonical Models have the following key applications:

   - Merger and acquisition—integrate applications with distinct data models.
   - Reducing application dependencies on an integrated data platform
   - Creating consistency in data nomenclature across business units
   - Development of a uniform data model
   - Assist with the Data Governance program.

3. A straightforward and practical analytical approach for assessing the caliber of COVID-19 datasets is Benford's law. As a result, it can be used to identify possible difficulties in huge data sets brought about by malicious a priori or a posteriori modification of datasets, as well as problems with rounding, data transfer, and treating observations to the limit of detection.

This model is offered to provide a method for determining how sound and relevant the data is from a quality standpoint for the desired data analysis goal.

7 Limitations

Although there are possible benefits of using the innovative method, there are also drawbacks. Due to the number of parameters optimized during data processing, the implementation of a data quality model necessitates a significant amount of time for data inspections and getting a comprehensive dataset. To accelerate computation time, a reasonable calculation time could be catered to use dynamic programming. Furthermore, the complexity of data sources is at the root of constraints in the accessibility of COVID-19 data and documentation.

8 Conclusions

This paper proposes data quality model to generate a set of actions to be taken so that to improve the quality of COVID-19 data. The framework is comprised of canonical data model, Four Adequacy levels, and Benford's law to manage and test the quality of the six CEMAC region countries COVID-19 data. The six CEMAC region countries COVID-19 data helped achieving an efficient DQM by reducing computing time and resources. The experiments we carried out on a big dataset demonstrated that the COVID-19 dataset's data quality can be restricted to a insignificant representative data sample. The end results are sets of generated adequacy levels based on data quality scores. Each adequacy level targets a data quality dimension for a dataset attribute. These proposed models are applied on the source dataset to strength and increase its quality. With the help of this methodology, we want to address the COVID-19 data issues discovered in six CEMAC region countries as a result of our expertise evaluating data quality models. As part of our ongoing research, we intend to create an automated optimization and model discovery system based on the outcomes of the data Quality dimension. Build a model and/or context metric for data quality dimension for big data and use it as a guide to generate data quality dimension metrics automatically.

Author contributions AN: investigation, methodology, conceptualization, data curation, formal analysis, validation, writing original draft, writing—review & editing. JZH conceptualization, methodology, conduct of the case studies, supervision, sources funding of the project, writing - review & editing, fundings. MK: validation, writing-review & editing. HW: validation, writing—review & editing.

Funding This paper was partially supported by National Natural Science Foundation China (NSFC) under Grant Number 61972261.

Availability of data and materials Datasets are available at: Our World In Data: COVID-19 Dataset https://ourworldindata.org/ or https://covid19.who.int/WHO-COVID-19-global-data.csv (Accessed on 15th June 2022).

Declarations

Conflict of interest No conflict of interest exists in the submission of this manuscript.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication All the authors listed have approved the manuscript for the publication.

References

1. Clowers A (2021) Covid-19: key insights from GAO's oversight of the federal public health response. Technical report, Government Accountability Office, Washington DC
2. Haq AU, Li JP, Ahmad S, Khan S, Alshara MA, Alotaibi RM (2021) Diagnostic approach for accurate diagnosis of covid-19 employing deep learning and transfer learning techniques through chest x-ray images clinical data in e-healthcare. Sensors 21(24):8219
3. Idrovo AJ, Manrique-Hernández EF (2020) Data quality of Chinese surveillance of covid-19: objective analysis based on who's situation reports. Asia Pac J Public Health 32(4):165–167
4. Ashofteh A, Bravo JM (2020) A study on the quality of novel coronavirus (covid-19) official datasets. Stat J IAOS 36(2):291–301
5. Barabesi L, Pratelli L (2020) On the generalized Benford law. Stat Probab Lett 160:108702
6. Alfaifi AA, Khan SG (2021) Utilizing data from twitter to explore the UX of Madrasati as a Saudi e-learning platform compelled by the pandemic. Arab Gulf J Sci Res 39(3):200–208
7. Khan S (2021) Visual data analysis and simulation prediction for covid-19 in Saudi Arabia using SEIR prediction model. Int J Online Biomed Eng 17(8):154–167
8. Else H (2020) Covid in papers: a torrent of science. Nature 588:553
9. Judson SD, Torimiro J, Pigott DM, Maima A, Mostafa A, Samy A, Rabinowitz P, Njabo K (2022) Covid-19 data reporting systems in Africa reveal insights for future pandemics. Epidemiol Infect 150:e119
10. Makoni M (2020) Covid-19 in Africa: half a year later. Lancet Infect Dis 20(10):1127
11. Khan S, Alfaifi A (2020) Modeling of coronavirus behavior to predict it's spread. Int J Adv Comput Sci Appl 11(5):394–399

12. Inzaule SC, Ondoa P, Loembe MM, Tebeje YK, Ouma AEO, Nkengasong JN (2021) Covid-19 and indirect health implications in Africa: impact, mitigation measures, and lessons learned for improved disease control. PLoS Med 18(6):1003666
13. Wolkewitz M, Puljak L (2020) Methodological challenges of analysing COVID-19 data during the pandemic. Springer, Berlin
14. Meehan MT, Rojas DP, Adekunle AI, Adegboye OA, Caldwell JM, Turek E, Williams BM, Marais BJ, Trauer JM, McBryde ES (2020) Modelling insights into the covid-19 pandemic. Paediatr Respir Rev 35:64–69
15. Apolloni B (2021) Inferring statistical trends of the covid19 pandemic from current data. Where probability meets fuzziness. Inf Sci 574:333–348
16. Stoto MA, Woolverton A, Kraemer J, Barlow P, Clarke M (2022) Covid-19 data are messy: analytic methods for rigorous impact analyses with imperfect data. Glob Health 18(1):1–8
17. Yazidi R, Aissi W, Bouguerra H, Nouira M, Kharroubi G, Maazaoui L, Zorraga M, Abdeddaiem N, Chlif S, El Moussi A et al (2019) Evaluation of the influenza-like illness surveillance system in Tunisia, 2012–2015. BMC Public Health 19(1):1–9
18. Visa TI, Ajumobi O, Bamgboye E, Ajayi I, Nguku P (2020) Evaluation of malaria surveillance system in Kano state, Nigeria, 2013–2016. Infect Dis Poverty 9(1):1–9
19. Cheng P, Gilchrist A, Robinson KM, Paul L (2009) The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. Health Inf Manag J 38(1):35–46
20. World Health Organization. Regional Office for the Eastern Mediterranean (2008) Strengthening National and Subnational Departments for Human Resources Development, vol 1. World Health Organization, Geneva
21. Gardner L, Ratcliff J, Dong E, Katz A (2021) A need for open public data standards and sharing in light of covid-19. Lancet Infect Dis 21(4):80
22. Ngueilbaye A, Wang H, Khan M, Mahamat DA (2021) Adoption of human metabolic processes as data quality based models. J Supercomput 77(2):1779–1817
23. Gualo F, Rodríguez M, Verdugo J, Caballero I, Piattini M (2021) Data quality certification using ISO/IEC 25012: industrial experiences. J Syst Softw 176:110938
24. Loshin D (2013) Big data analytics: from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph. Elsevier, Amsterdam
25. Ngueilbaye A, Wang H, Mahamat DA, Elgendy IA, Junaidu SB (2021) Methods for detecting and correcting contextual data quality problems. Intell Data Anal 25(4):763–787
26. Caballero I, Gualo F, Rodríguez M, Piattini M (2022) BR4DQ: a methodology for grouping business rules for data quality evaluation. Inf Syst 109:102058
27. Khan S, Alqahtani S (2020) Big data application and its impact on education. Int J Emerg Technol Learn (iJET) 15(17):36–46
28. Merino J, Caballero I, Rivas B, Serrano M, Piattini M (2016) A data quality in use model for big data. Futur Gener Comput Syst 63:123–130
29. Ding X, Wang H, Su J, Wang M, Li J, Gao H (2020) Leveraging currency for repairing inconsistent and incomplete data. IEEE Trans Knowl Data Eng 34:1288–302
30. Mooney SJ, Pejaver V (2018) Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Health 39:95
31. Hussein AA (2020) Fifty-six big data V's characteristics and proposed strategies to overcome security and privacy challenges (BD2). J Inf Secur 11(4):304–328
32. Ivorra B, Ferrández MR, Vela-Pérez M, Ramos AM (2020) Mathematical modeling of the spread of the coronavirus disease 2019 (covid-19) taking into account the undetected infections. The case of china. Commun Nonlinear Sci Numer Simul 88:105303
33. Krishna MV (2020) Mathematical modelling on diffusion and control of covid-19. Infect Dis Model 5:588–597
34. Strong DM, Lee YW, Wang RY (1997) Data quality in context. Commun ACM 40(5):103–110
35. Carleson L (1966) On convergence and growth of partial sums of Fourier series. Acta Math 116:135–157
36. Kakutani S (1941) A generalization of Brouwer's fixed point theorem. Duke Math J 8:457–459
37. Lenzerini M (2002) Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, pp 233–246
38. Koch AJ (2001) Factors influencing market and entry mode selection: developing the mems model. Mark Intell Plan 19:351–361

39. Friedman MT, Levy AY, Millstein TD et al (1999) Navigational plans for data integration. AAAI/IAAI 1999:67–73
40. Miller R, Haas L, Hernandez M (2000) Schema mapping as query discovery. In: 26th VLDB, Cairo
41. Benford F (1938) The law of anomalous numbers. Proc Am Philos Soc 78:551–572
42. Pietronero L, Tosatti E, Tosatti V, Vespignani A (2001) Explaining the uneven distribution of numbers in nature: the laws of Benford and ZIPF. Physica A 293(1–2):297–304
43. Graham RL, Knuth DE, Patashnik O, Liu S (1989) Concrete mathematics: a foundation for computer science. Comput Phys 3(5):106–107
44. Berger A, Hill TP (2015) An introduction to Benford's law. Princeton University Press, Princeton
45. Ngueilbaye A, Wang H, Mahamat DA, Junaidu SB (2021) Modulo 9 model-based learning for missing data imputation. Appl Soft Comput 103:107167
46. Tuli S, Tuli S, Tuli R, Gill SS (2020) Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing. Internet Things 11:100222

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Alladoumbaye Ngueilbaye[1,2] · Joshua Zhexue Huang[1,2] · Mehak Khan[3] · Hongzhi Wang[4]

✉ Alladoumbaye Ngueilbaye
angueilbaye@szu.edu.cn

Joshua Zhexue Huang
zx.huang@szu.edu.cn

Mehak Khan
mehakkha@oslomet.no

Hongzhi Wang
wangzh@hit.edu.cn

[1] Big Data Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, Guangdong, China

[2] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, Guangdong, China

[3] Department of Computer Science, AI Lab, Oslo Metropolitan University, Oslo, Norway

[4] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China