Analysis of the different sources of secondary data in Peru

1. **Objective:**
   To evaluate the quality and use of epidemiological information produced by different information systems in Peru.

2. **Diseases of interest**
   The evaluation of data sources will be made on the basis of the following diseases:
   - Dengue
   - Malaria
   - Rabies
   - Leishmaniasis
   - COVID-19
   - Leptospirosis
   - Acute Respiratory Infections (ARI)

3. **Source of data**
   The secondary data sources used to evaluate all diseases except covid-19 were as follows:

   a. ***NOTI-WEB***: is the system that collects data from notifiable diseases, managed by CDC-Peru. We may not access to the database, but CDC Peru offered us calculate indicators requested. The fields contained in that system, that are common to all diseases of interest in our study are:
      - Gender
      - Age
      - Place of infection
      - Date of notification
      - Type of diagnosis (confirmed, probable, discarded)
      - Date of onset symptoms

   b. ***HIS*** (Health information System): is the outpatient information system. Fields shared with us by MINSA are:
      - Place of residence (department/region, province, district)
      - Date of visit
      - Age
      - Gender

      The request for information on the diseases to be analyzed was made according to the CIE-10 codes included in the *"List of diseases and events subject to epidemiological surveillance in Peru"* of the Sanitary Directive N°046-MINSA/DGE-V.01. The database received contains no missing data.

   c. ***SINADEF***: is the electronic death certificate system implemented in 2017. The fields included in the open access database are:
      - Gender
      - Age
      - Place of residence (department/region, province, district)
      - Date of death

- Marital status
- Education level
- Up to six ICD 10 codes (four direct causes and two indirect)
- Short description of causes of death (four direct causes and two indirect)
- Other fields.

To COVID-19

The data sources for covid-19, which are open databases, were the following:

Data sources	Description	Variables
Covid-19 deaths dataset	This is the daily record of deaths from Covid-19. Each record is equal to one person.	Age Sex Place of residence (department/region, province, district)
PCR test dataset	This is the dataset of molecular tests registered in the Information System of the National Network of Public Health Laboratories in Peru (NETLAB). This system is managed by the National Institute of Health.	Age Sex Date of sample collection Type of sample Sample result Institution that took the test Place of the institution (department/region, province, Diresa, Red)
Positive results dataset	It contains daily information on covid-19 positive cases confirmed with any type of test and presenting symptoms.	Sex Age Place of residence (department/region, province, district) Type of test (PCR, test rapid, antigen test) Date of test result
Covid-19 Suspect (triage)	This database contains information on the person suspected of Covid-19	Date of contact of the person (via phone call 113) Date of symptom onset Presence of symptoms: fever, cough, headache, etc.
Clinical attention	Database containing the patient's covid-19 status, admission, evolution, death, discharge and referral.	Date of admission to health facility Date of discharge Date of death Date of reference Destination establishment Others fields
Hospitalizations	This database contains information on hospitalized persons, linking information on vaccine doses and death by covid-19.	Date of admission to the hospital. Date of admission to ICU Date of admission to Intermediate Care Use supplemental oxygen Use mechanical ventilation Evolution in the hospital Date of vaccine dose Vaccine brand Date of death by COVID Place of residence

4. **Inclusion criteria:**

For the HIS and Notiweb databases all registered cases, minors or adults, were analyzed. For ARIs, only children under 5 years of age were included, since they are the target population for epidemiological surveillance in Peru. In addition, records from 2011 to 2022 were included. For SINADEF, cases registered from 2017 to 2022 were analyzed.

To analyze the databases for covid-19, all cases from 2020 to 2022 were included. Data will be downloaded from the Peruvian Open Data portal.

5. **Attributes to be evaluated**

The indicators that were constructed to evaluate the quality of the information systems for all diseases except covid-19 are shown below.

#	Indicator definition	Numerator and denominator	Level	Data source
A1. Completeness				
1a	**Proportion of cases in a database with no missing required information**	**Numerator:** number of deaths with complete information on age, gender, location. **Denominator:** Total deaths recorded.	Year Region Disease	**SINADEF**
1b	**Proportion of missing information required by field (age, gender, location)**	**Numerator:** number of deaths with complete information for each of the three fields **Denominator:** Total deaths recorded	Year Disease (for each field)	**SINADEF**
A2. Validity				
2a	**Proportion of coding errors within a dataset**	**Numerator:** Number of deaths with the specific disease described as a cause of death in any of the six fields and a compatible ICD10. **Denominator:** Number of deaths with the specific disease described as a cause of death in any of the six fields	Year Region Disease	**SINADEF**
A3. Sensitivity				
3a	**Sensitivity of surveillance system (NOTI)**	*Outpatient Information* **Numerator**: Number of confirmed cases in NOTI **Denominator**: Number of cases registered in HIS	Year Region Disease	**NOTI** **HIS**
A4. Positive Predictive Value				
4a	**Proportion of confirmed cases reported through the surveillance system in case-based surveillance**	**Numerator:** Number of confirmed cases. **Denominator:** Number of confirmed plus discarded cases	Year Region Disease	**NOTI**
A5. Timeliness				
5a	**Main time delays**	**Time to report:** Date reported – date onset	Year Region Disease	**NOTI**
A6. Representativity				
6a	**Ratio of the number of districts with notifications in surveillance system to the number of districts with cases in HIS**	**Numerator:** number of districts with notification in Noti **Denominator:** number of districts with registered in HIS	Year Disease	**NOTI** **HIS**

For covid-19, the indicators that were constructed were as follows

#	Indicator definition	Numerator and denominator	Level	Data source
A1. Completeness				
1a	**Proportion of cases recorded in a database with no missing required information**	**Numerator:** total cases recorded with no missing information **Denominator:** total cases recorded, including unknown and missing items	Year Region Sex	**PCR** **Positives** **Deaths**
A2. Validity				
2a	**Proportion of inconsistencies and errors within a dataset (inconsistent dates or out of range)**	**Numerator:** total cases with dates out of the range of the pandemic (less than "March 2020" or greater than dataset date). **Denominator:** total cases in the dataset	Sex Institution	**PCR**
A3. Concordance				
3a	**PCR testing results concordance** Refers to the proportion of concordance in positive PCR tests between two different data sources	**Numerator**: Number of positive PCR results in PCR dataset **Denominator**: Number of positive PCR results in positive dataset	Year Region Sex	**PCR** **Positives**
A4. Timeliness				
4a	**Main time delays**	**Delay to report:** date of report (call) – date of onset (date of symptoms) **Delay to lab result:** date of lab result – date of onset of symptoms **Delay to attention:** date of admission to hospital– date of onset of symptoms	Year Region Sex	**Positives** **Triage** **Hospitalization**

6. Statistical methods and statistical package:

Data will be analyzed in R-Studio. Frequencies and proportions were calculated according to the previously mentioned indicators and for the timeliness indicators, means, medians and interquartile ranges were calculated.

The indicators were calculated according to year, region, diseases. Some indicators were calculated according to sex and care institution. In addition, trend graphs of the indicators will be constructed.