DATA CLEANING TIP SHEET

mHero's success relies on having accurate information, or data, about health workers, which is pulled from their human resource records. Errors in the data limit its use. If the data used for mHero messaging contains incorrect phone numbers, messages cannot be sent to their intended recipients. If health worker cadres are not correctly assigned, information may be requested from the wrong group of providers, impeding data analysis.

Data cleaning, which can be done for all types of datasets, is an ongoing process of identifying and correcting data that is incomplete, inaccurate, or otherwise problematic. If data cleaning is neglected, errors can accumulate over time. In the case of mHero, this can result in the failure of important messaging campaigns, which can cause stakeholders to lose confidence in the platform.

High-quality data that is fit for use is data that is accurate, consistent, valid, complete, uniform, traceable, and timely.

Ensure **accuracy** of source data.

- In many cases, the source system for health worker information will be iHRIS or another HRIS.

- There may be different sources for different types of data, such as facility records.

- Avoid using RapidPro as the source of data.

- It may be necessary to do a thorough initial cleaning to ensure that the data in your source system is accurate.

Maintain **consistency** of records.

- When two or more systems are in use, data may become out of sync.

- Designate one system as the primary source of data for mHero (also called the "source of truth" system). One way to do this is by reviewing metadata associated with the records to determine the system in which data was most recently changed.

- Develop procedures to ensure that this "source of truth" system is the only point of data entry.

Validate data.

- Data should meet specific rules or conditions. Phone numbers, for example, should have the same number of digits and no characters other than digits.

- Know whether country codes are required: it is common for country codes to be part of the phone number in some systems and not in others.

- Add validation mechanisms to data entry processes to enforce these rules if they are not already in place.

Make sure information is **complete.**

- Collect unknown information and enter or import it into the designated source system.

- You may have to use external data sources, such as lists maintained by professional associations, to find missing phone numbers or other information.

Follow **uniform** standards.

- Data used to identify groups need to be labeled the same across systems, such as a predefined list of cadre names.

- Different systems may format the same information differently. In those cases, you need to define a method to standardize the dataset before using it in mHero.

Make information **traceable.**

- Document the source of all data that are added or updated. Later users should be able to trace data sources easily.

Update information to keep it **timely.**

- All health worker information, particularly phone numbers, can change often.

- Ensure there is a method for keeping information updated. For example, human resources officers could develop a procedure to regularly confirm employees' phone numbers.

Plan for data cleaning. Be proactive about cleaning data by making it a regular activity. Regular data cleaning reduces the workload burden and ensures that the data mHero needs are always fit for use.

Know what to check for. The value of the mHero platform is using information about health workers to send them targeted text messages. Therefore, you need to ensure that the data fields used to target messages are regularly validated or updated and free of errors:

- First and last names

- Phone numbers

- Facilities of employment

- Cadres

- Geographic locations, such as home address, district or county, etc.

Of these, phone number data can present the most problems. Phone numbers are often out of date or missing altogether. The same phone number may be associated with multiple records, or one person may have multiple phone numbers. Your strategy should include verification of the primary phone number and aim to remove duplicates.

Have a strategy. Your plan for data cleaning and maintenance should specify how often the data will be audited, verified, and updated and who is responsible for those tasks. When considering frequency, think about how quickly critical information, like phone numbers, may fall out of date.

It may be necessary to verify the data more than just once a year to keep up with changes.

Determine an approach for auditing the data and updating records. You may choose to spot-test randomly selected fields or run scripts to identify issues, but it will be necessary to do some manual checking to make sure you catch and address all errors. The procedure may include contacting health workers directly or working with human resources officers to collect updated information. You can use mHero to send messages to health workers to confirm the information entered into their records.

Document your decisions in your data-cleaning strategy:

- "Source of truth" system

- Rules or conditions for specific data fields

- Standards used

- Frequency of auditing, verifying, and updating data

- Data maintenance procedures, roles, and responsibilities.

A proactive strategy for maintaining good data hygiene will reduce how much time data cleaning takes. Having clean, error-free data that is fit for use will expand the functionality and utility of the platform, boost confidence in mHero, and bolster its support.

CONTACT
digitalhealth@intrahealth.org
www.mhero.org

mHero