Getting and Cleaning Data

CIHR Course Week 4

Bethany Hedt-Gauthier

Izzie Fulcher

Teaching Objectives

- Overview
- Outliers
- Missing data

Overview

Health management information systems

- Countries have always generated/used data.

- Nationalized systems garnered global attention beginning in mid-1990s.
  - Standardize processes
  - Common indicators
  - Universal systems such as DHIS2

2. Expectations of a country health information system

Health information systems serve multiple users and a wide array of purposes that can be summarized as the generation of information to enable decision-makers at all levels of the health system to identify problems and needs, make evidence-based decisions on health policy and allocate scarce resources optimally.[2] Data from different sources are used for multiple purposes at different levels of the health care system.

- **Individual** level data about the patient's profile, health care needs, and treatment serve as the basis for clinical decision-making. Health care records provide the basis for sound individual clinical care. Problems can arise when health workers are overburdened by excessive data and reporting demands from multiple and poorly coordinated subsystems.
- **Health facility** level data, both from aggregated facility-level records and from administrative sources such as drug procurement records, enable health care managers to determine resource needs, guide purchasing decisions for drugs, equipment and supplies, and develop community outreach. Data from health facilities can provide immediate and ongoing information relevant to public health decision-making but only if certain conditions are met. The data must be of high quality, relate to all facilities (public and private), and be representative of the services available to the population as a whole.
- **Population** level data are essential for public health decision-making and generate information not only about those who use the services but also, crucially, about those who do not use them. Household surveys have become a primary source of data in developing countries where facility-based statistics are of limited quality. But household surveys are needed everywhere because they are the only good source of information on individual beliefs, behaviours and practices that are critical determinants of health care use and of health status.
- **Public health surveillance** brings together information from both facilities and communities with a focus mainly on defining problems and providing a timely basis for action. This is especially so when responses need to be urgent, as in the case of epidemic diseases. The need for timeliness of reporting and response, and the requirement for effective linkages to those in authority with the responsibility for disease control, impose additional requirements on health information systems.

Go to **www.menti.com** and use the code **9859 3980**

What is your perspective on the HMIS from your country (or the country you support)?

Mentimeter

PubMed.gov

"health management information systems"

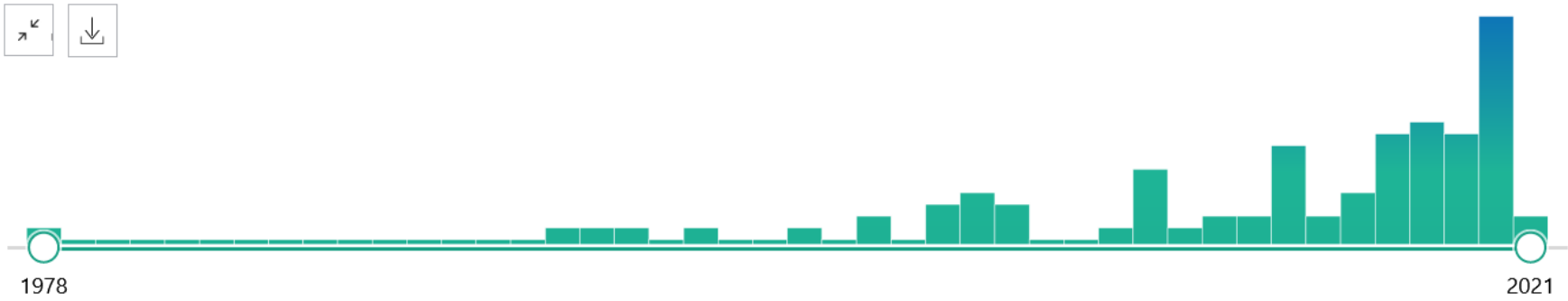
Search

Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Best match Display options

RESULTS BY YEAR

90 results



PubMed.gov

DHIS2

Search

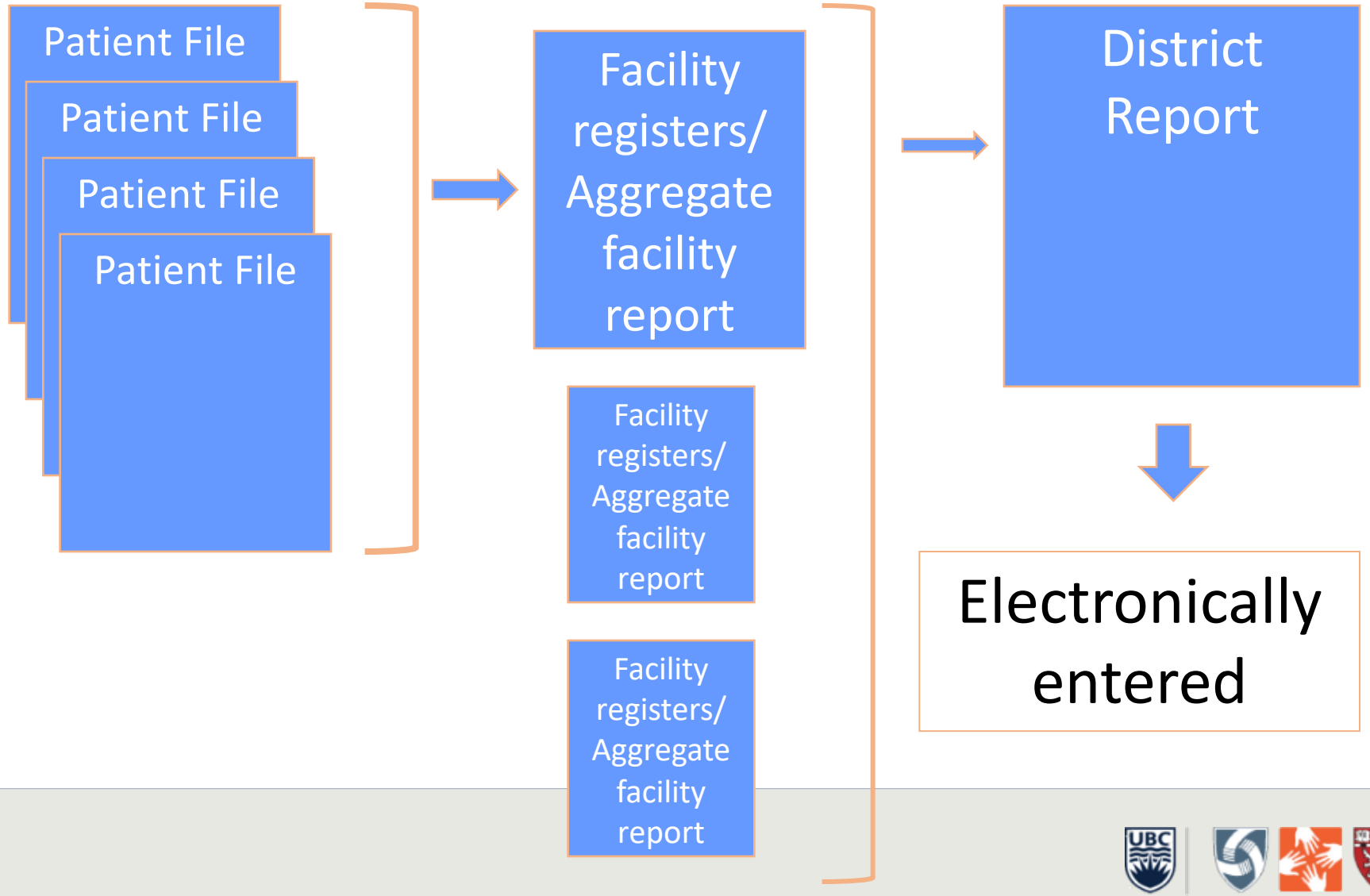
Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Best match Display options

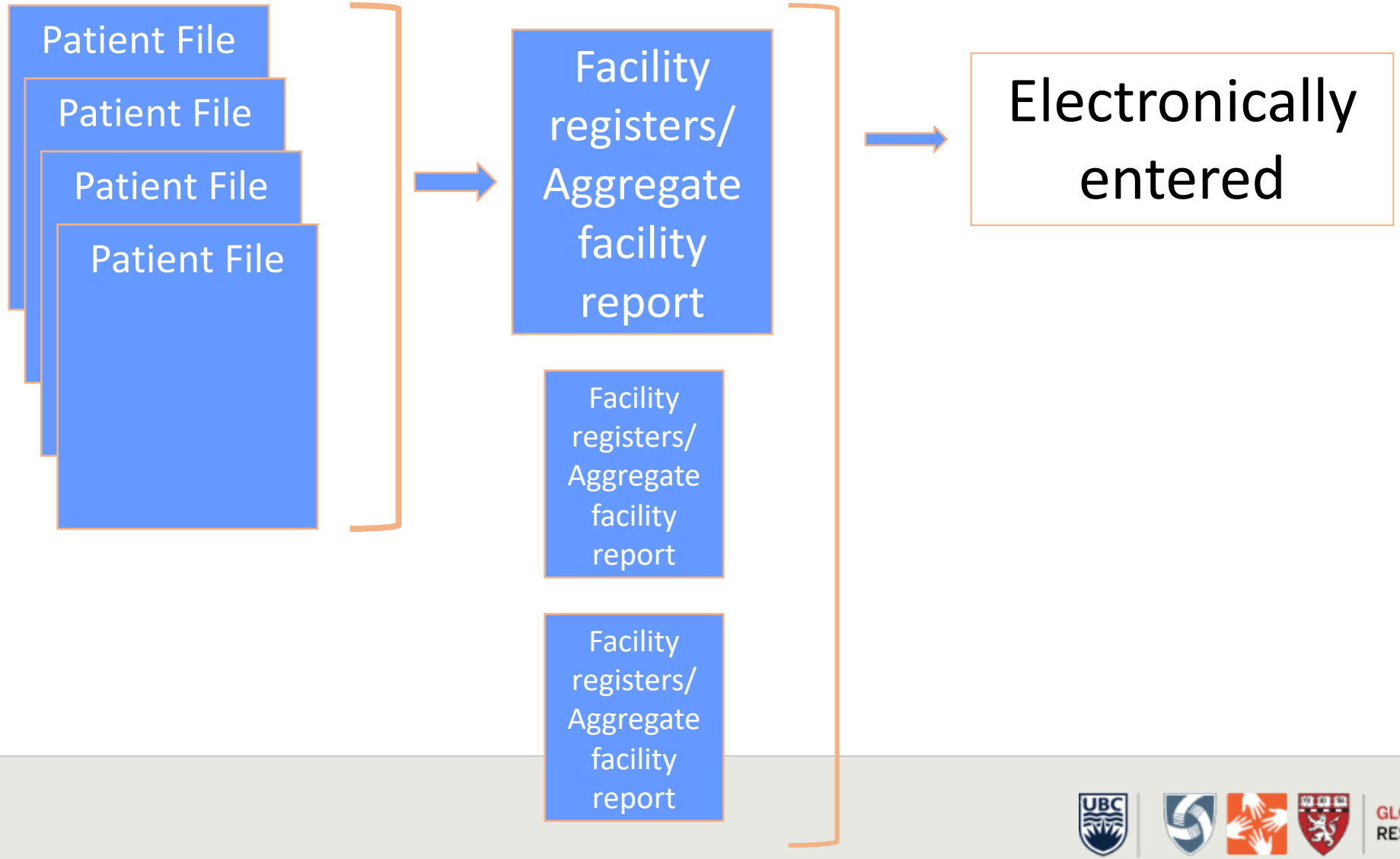
RESULTS BY YEAR

72 results

Flow of HMIS data



Flow of HMIS data



Quality of data recorded?

- Is it complete?
- Is it valid?
- Is it reliable?
- Is it accurate?

Completeness and validation

Global Health Action

SHORT COMMUNICATION

Toward utilization of data for program management and evaluation: quality assessment of five years of health management information system data in Rwanda

Marie Paul Nisingizwe[1]*, Hari S. Iyer[1,2], Modeste Gashayija[3], Lisa R. Hirschhorn[2,4,5], Cheryl Amoroso[1], Randy Wilson[3,6], Eric Rubyutsa[3], Eric Gaju[3], Paulin Basinga[7], Andrew Muhire[3], Agnès Binagwaho[3,5,8] and Bethany Hedt-Gauthier[1,5,9]

[1]Research Department, Partners In Health/Inshuti Mu Buzima, Kigali, Rwanda; [2]Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA, USA; [3]HMIS Department, Ministry of Health, Kigali,

Table 2. Completeness of facility reporting and indicator data (2008–2012)

	2008 (%)	2009 (%)	2010 (%)	2011 (%)	2012 (%)
National district completeness rate	95	99	98	100	100
Districts with completeness rate below 80%	7	0	0	0	0
Completeness of indicator data	88	91	89	90	95
Proportion of district with more than 20% missing values	7	0	3	3	0

Table 3. Outliers and internal consistency between indicators (2008–2012)

	2008	2009	2010	2011	2012
Extreme and moderate outliers					
Proportion of values that are moderate outliers[a]	0%	0%	0%	0%	0%
Proportion of values that are extreme outliers[a]	0%	0%	0%	0%	0%
Internal consistency between DTP1 and ANC1					
National DTP1/ANC1 ratio	0.87	0.97	0.87	0.90	0.94
Proportion of districts with DTP1/ANC1 ratio 33% above national ratio	10%	3%	0%	0%	0%
Proportion of districts with DTP1/ANC1 ratio 33% below national ratio	0%	13%	0%	0%	0%
Internal consistency between DTP1 and DTP3					
National DPT3/DTP1 ratio	0.96	1.00	1.01	0.97	0.99
Proportion of districts where DTP3 is 2% greater than DTP1	13%	17%	3%	0%	23%

Reliability

PLOS ONE

RESEARCH ARTICLE

Health management information sys
(HMIS) data verification: A case study
districts in Rwanda

Alphonse Nshimyiryo[1]*, Catherine M. Kirk[1], Sara M. Sauer[2], Emmanuel Ntawuyirusha[3], Andrew Muhire[3], Felix Sayinzoga[4], Bethany

1 Maternal and Child Health Program, Partners In Health/Inshuti Mu Buzima, Kigali, R
of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States
3 Planning, Health Financing and Information Systems, Ministry of Health, Kigali, Rwa

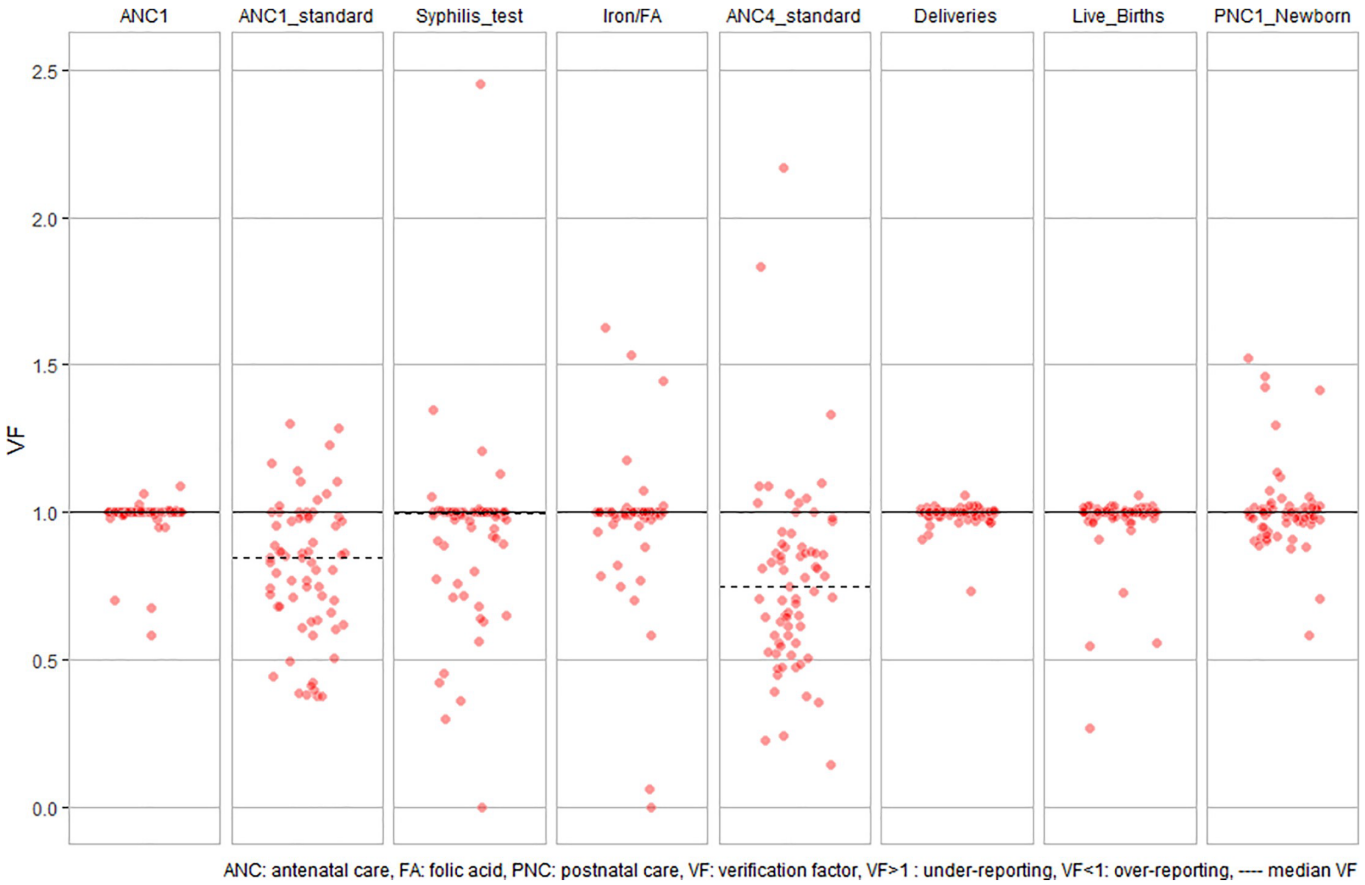


Fig 1. Facility-level verification factors (VF) by data element. The median is not visible when it is very close to or equal to 1.

Accuracy

J Community Health (2015) 40:625–632
DOI 10.1007/s10900-014-9977-9

ORIGINAL PAPER

Data for Program Management: An Accuracy As Collected in Household Registers by Community in Southern Kayonza, Rwanda

Tisha Mitsunaga · Bethany L. Hedt-Gauthier · Elias Ngizwenayo · Didi Bertrand Farmer · Erick Gaju · Peter Drobac · Paulin Basinga · Lisa Hirschhorn · Michael L. Rich · Peter J. Winch · Fidele Ngabo · Cathy Mugeni

Published online: 11 December 2014

Abstract Community health workers (CHWs) collect data for routine services, surveys and research in their widely under-report planning method (

Table 3 Point estimates and 95 % CIs of data accuracy comparing household registers to interview by indicator and health center/district

Health center/district	Indicator					
	Number of children under-five	Number of women 15–49	Number of women on modern family planning method	Type of family planning method	Number of home deliveries	Composite
	Point estimate (95 % CI)					
HC1	91.3 % (86.9 %, 95.8 %)	79.5 % (73.2 %, 85.8 %)	88.0 % (83.0 %, 92.9 %)	83.8 % (78.2 %, 89.4 %)	100.0 % (NA)	61.5 % (54.1 %, 68.9 %)
HC2	91.3 % (86.3 %, 96.2 %)	87.4 % (81.8 %, 93.1 %)	86.8 % (80.9 %, 92.7 %)	85.3 % (79.1 %, 91.5 %)	98.3 % (95.9 %, 101 %)	68.6 % (60.8 %, 76.5 %)
HC3	86.6 % (81.5 %, 91.8 %)	91.6 % (87.5 %, 95.7 %)	91.9 % (88.2 %, 95.6 %)	86.3 % (81.2 %, 91.3 %)	99.5 % (98.5 %, 100 %)	68.5 % (61.6 %, 75.4 %)
HC4	88.5 % (83.2 %, 93.7 %)	86.9 % (81.4 %, 92.4 %)	86.5 % (80.9 %, 92.1 %)	85.3 % (79.6 %, 90.9 %)	98.5 % (96.5 %, 101 %)	61.1 % (53.2 %, 69.1 %)
HC5	84.5 % (78.6 %, 90.4 %)	84.4 % (78.4 %, 90.4 %)	87.0 % (81.6 %, 92.5 %)	86.9 % (81.4 %, 92.4 %)	99.4 % (98.4 %, 101 %)	64.7 % (56.8 %, 72.5 %)
HC6	89.4 % (84.3 %, 94.4 %)	90.6 % (85.8 %, 95.3 %)	88.5 % (83.4 %, 93.7 %)	87.7 % (82.3 %, 93.0 %)	100.0 % (NA)	71.7 % (64.3 %, 79.0 %)
HC7	83.3 % (71.2 %, 95.5 %)	91.7 % (82.6 %, 101 %)	88.9 % (78.6 %, 99.1 %)	88.9 % (78.6 %, 99.1 %)	100.0 % (NA)	66.7 % (51.3 %, 82.0 %)
HC8	87.3 % (83.3 %, 91.4 %)	90.0 % (86.4 %, 93.6 %)	89.5 % (85.8 %, 93.3 %)	87.4 % (83.3 %, 91.5 %)	99.7 % (99.2 %, 100 %)	70.4 % (64.9 %, 76.0 %)
District	88.3 % (86.4 %, 90.2 %)	86.8 % (84.7 %, 88.8 %)	88.2 % (86.3 %, 90.2 %)	86.1 % (84.0 %, 88.2 %)	99.4 % (98.9 %, 99.9 %)	66.5 % (63.7 %, 69.3 %)

Table 4 Percentage over-reporting, median difference and IQR of discordant household register entries compared to interview by indicator and health center/district

Health center/district	Indicator					
	% Household register > Household visit | Median difference (IQR)					
	Number of children under five years		Number of women 15–49 years		Number of women on modern family planning method	
HC1	12.3 %	−1 (−1,−1)	66.7 %	1 (−1,1)	20.8 %	−1 (−1,−1)
HC2	48.5 %	0 (−1,1)	50.2 %	−1 (−1,1)	8.8 %	−1 (−1,−1)
HC3	30.7 %	−1 (−1,1)	76.5 %	1 (−1,1)	16.4 %	−1 (−1,−1)
HC4	25.7 %	−1 (−1,0)	67.2 %	1 (−1,1)	22.5 %	−1 (−1,−1)
HC5	38.5 %	−1 (−2,1)	53.7 %	1 (−1,1)	19.6 %	−1 (−1,−1)
HC6	27.1 %	−1 (−1,1)	60.0 %	1 (−1,1)	12.1 %	−1 (−1,−1)
HC7	33.3 %	−1 (−2,1)	33.3 %	−1 (−1,1)	25.0 %	−1 (−1, 0)
HC8	44.1 %	−1 (−1,1)	35.3 %	−1 (−1,1)	35.9 %	−1 (−1,1)
District	33.7 %	−1 (−1,1)	57.4 %	1 (−1,1)	19.7 %	−1 (−1,−1)

High-level thoughts on HMIS quality

- There is no such thing as perfect data.

- Important to monitor quality of your data:
  - Good enough to use?
  - Ways to improve?

- Even imperfect data can be usable data.
  - Can throw out clear mistakes.
  - If bias is consistent, then can still detect outliers.
  - May be able to fill in missing values.

Outliers

A note

In this session, we will discuss outliers and missing data in the context of fitting a time series model for a single facility and indicator

Types of outliers

- **Outliers are data points that are very different from the majority of observations in the time series**

- **There are various types of outliers:**
  - **Errors in data entry**
  - **Unusual cases**
    - Global outlier: raw values are significantly larger or smaller than the rest of the values.
    - Contextual outlier: higher or lower value than you would expect based on the patterns and trends of the time series

How do we detect outliers?

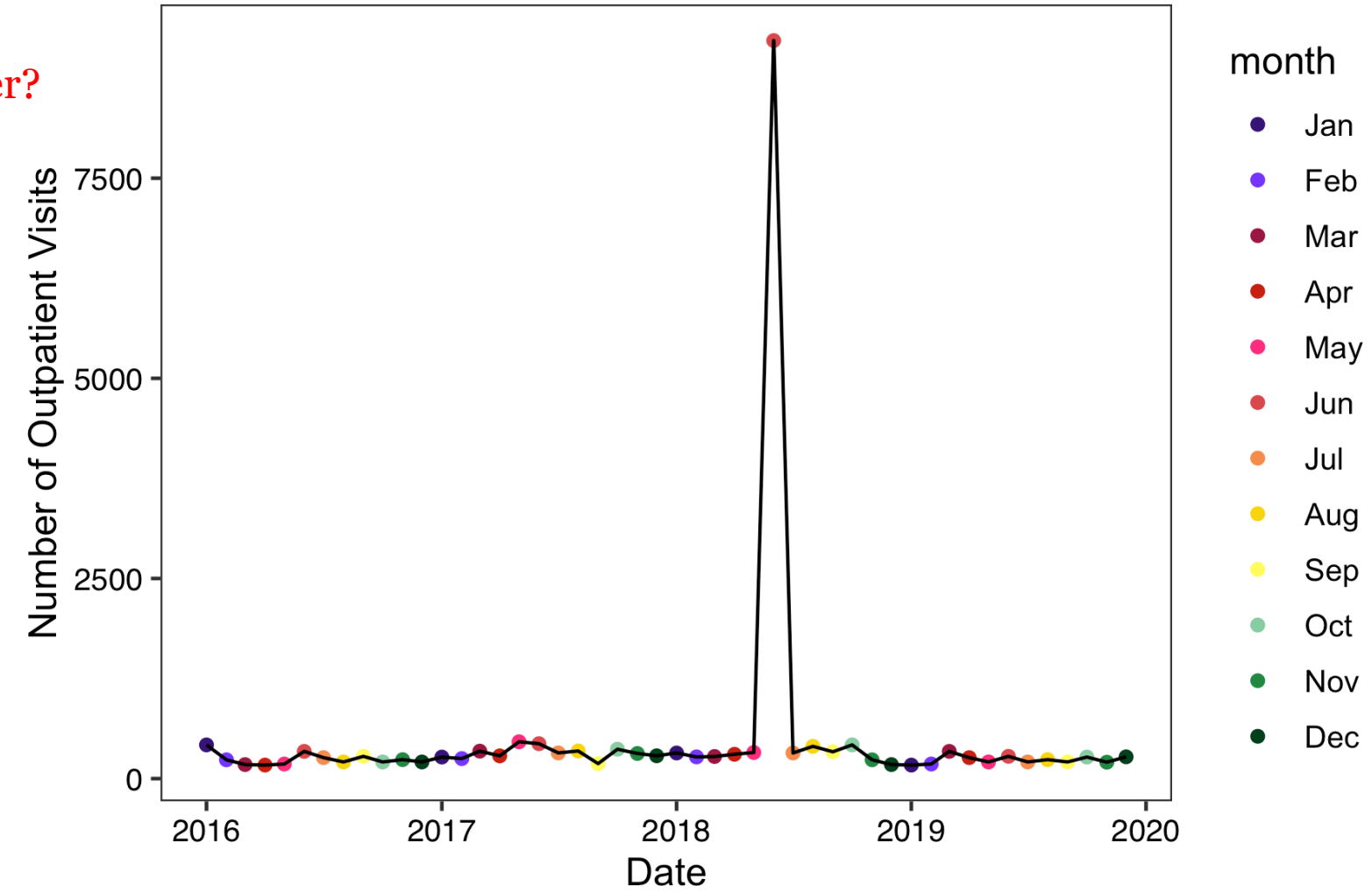
- **Method 1: Visual inspection**
  - Hard to detect outliers with time series data due to seasonality

- **Method 2: Tukey's rule** (larger than 1.5*IQR values)
  - Does not capture the seasonal nature
  - Boxplots by month – but would not capture trend & need many years of data
  - Boxplots by year – but would not capture seasonality

- **Method 3: Fit time series model and apply statistical test**
  - Captures trends, but it is model dependent

Method 1: Visual inspection

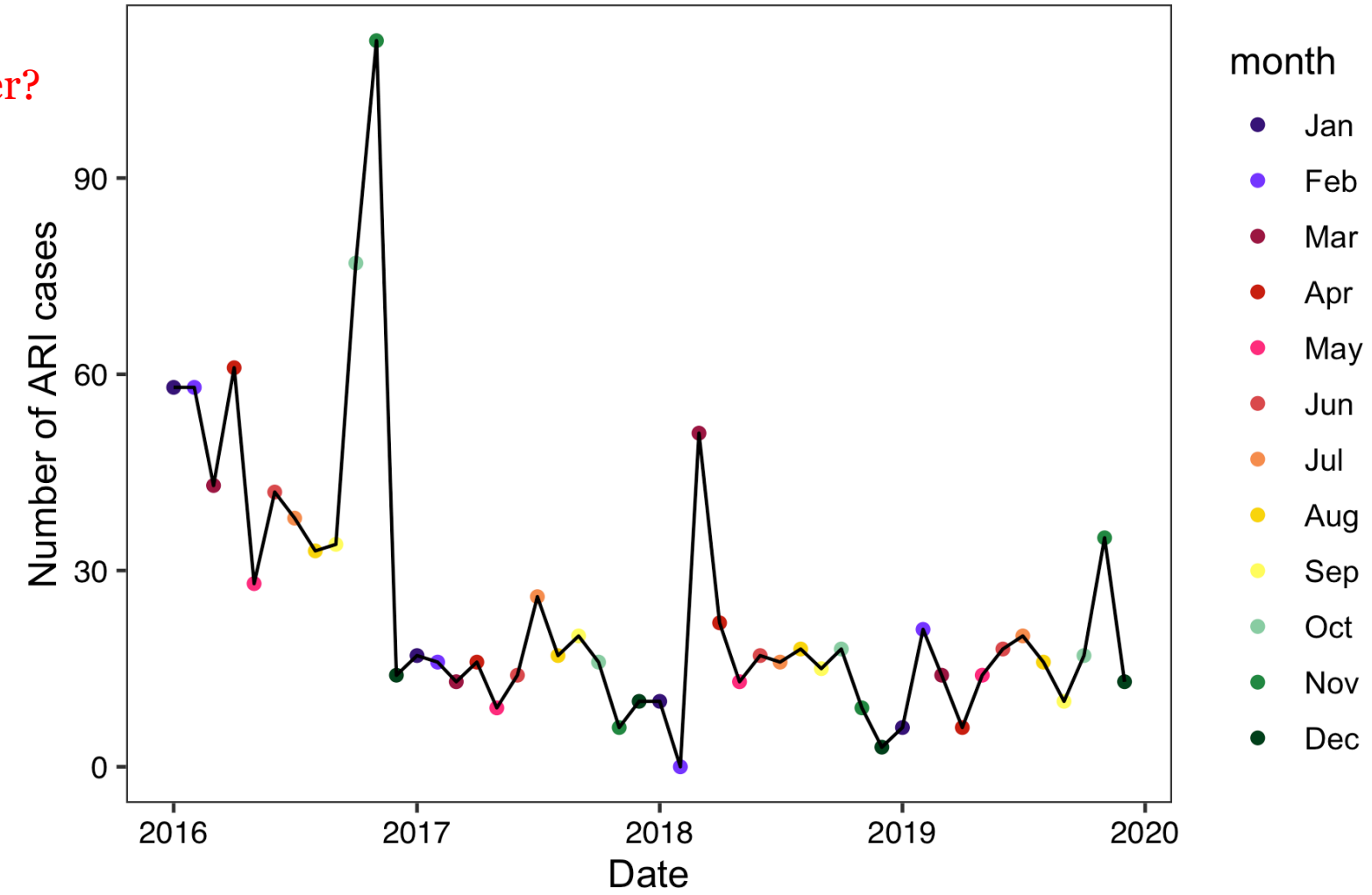
- Error?
- Global outlier?
- Contextual outlier?



Jayproken Clinic in Liberia

Method 1: Visual inspection

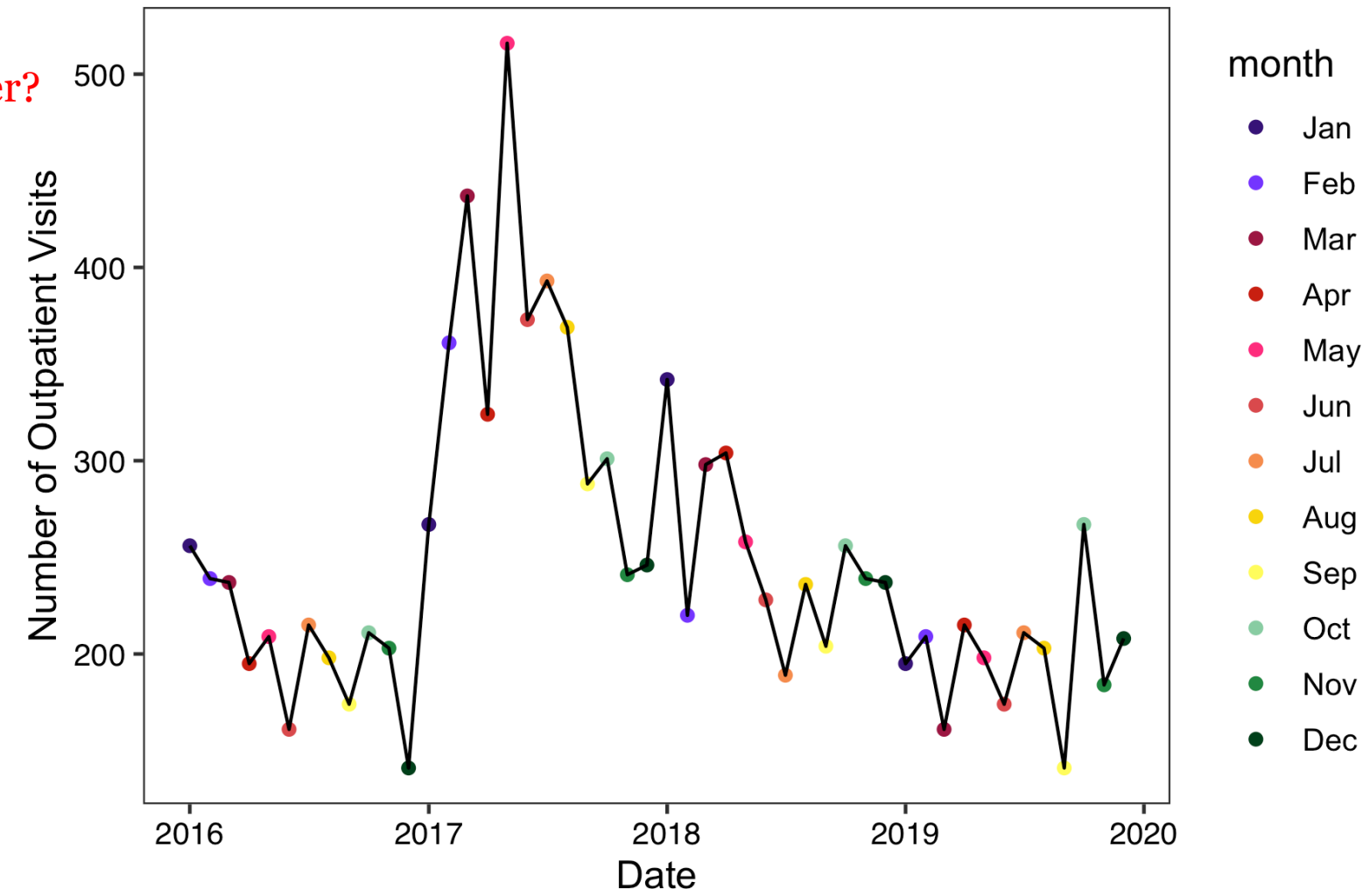
- Error?
- Global outlier?
- Contextual outlier?



St. Francis Clinic in Liberia

Method 1: Visual inspection

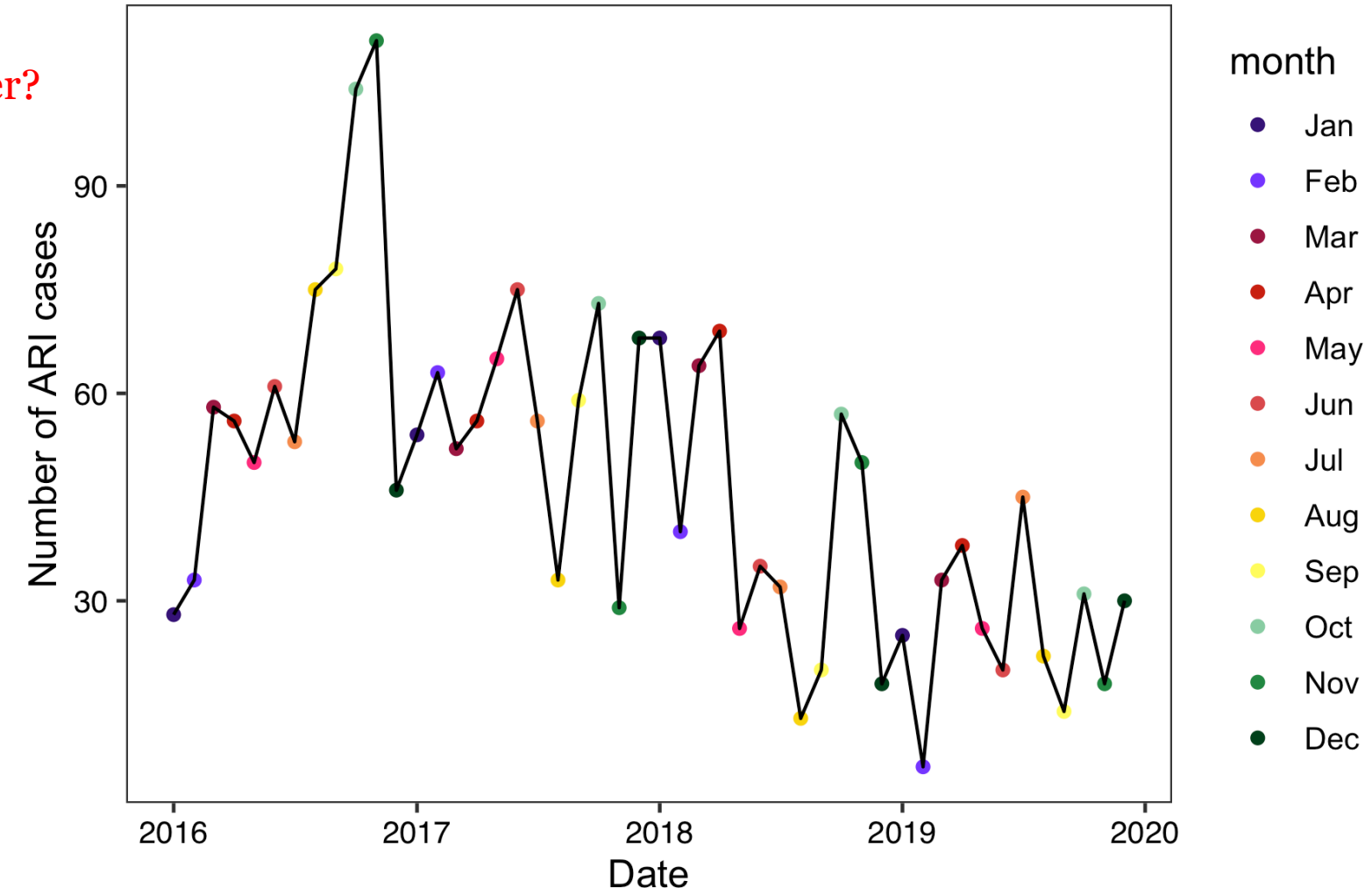
- Error?
- Global outlier?
- Contextual outlier?



Yangaya Clinic in Liberia

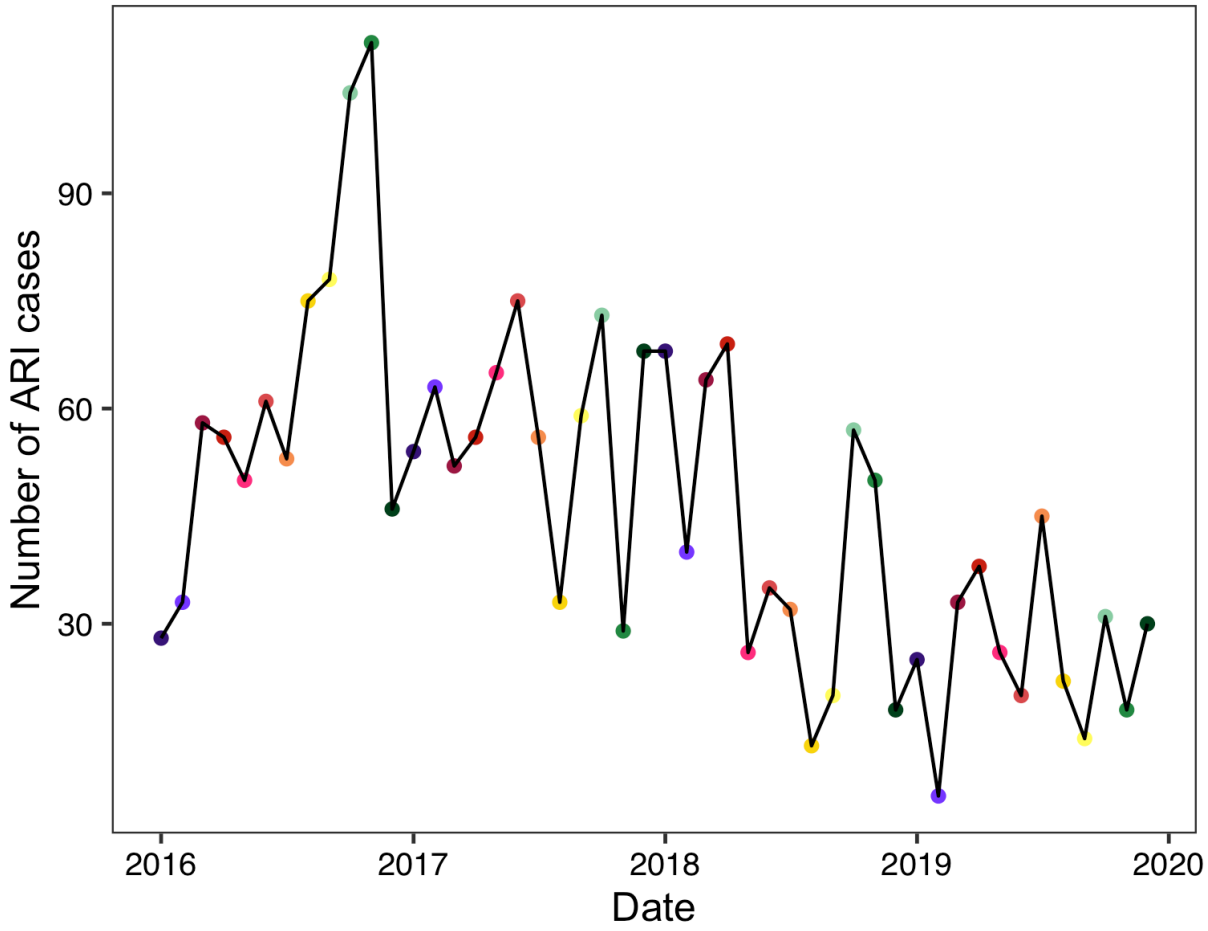
Method 1: Visual inspection

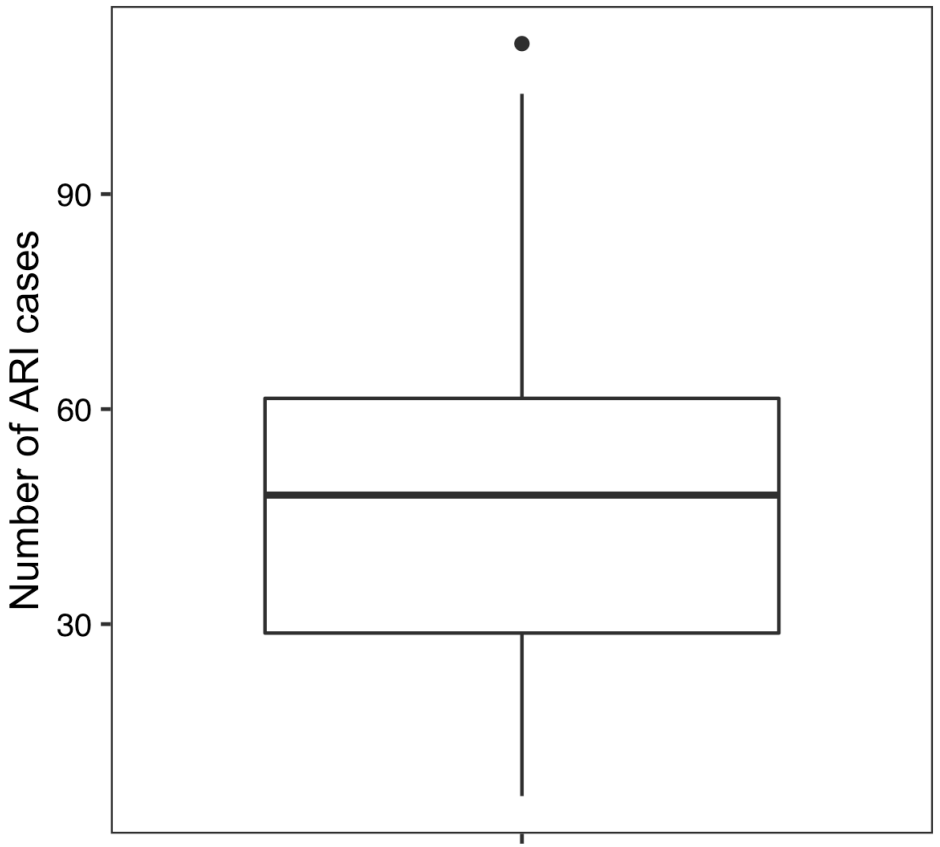
- Error?
- Global outlier?
- Contextual outlier?



Boniken Clinic in Liberia

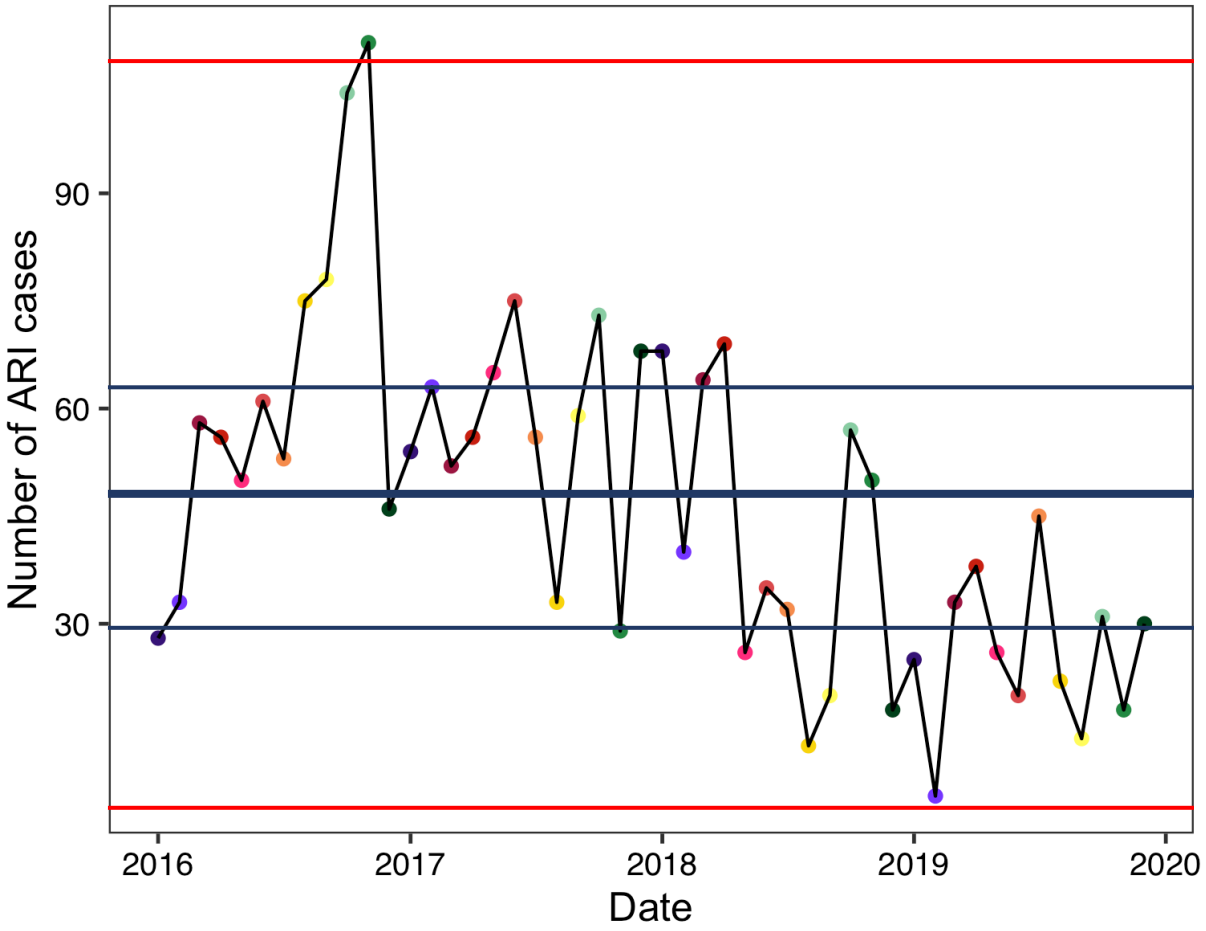
Method 2: Tukey's rule

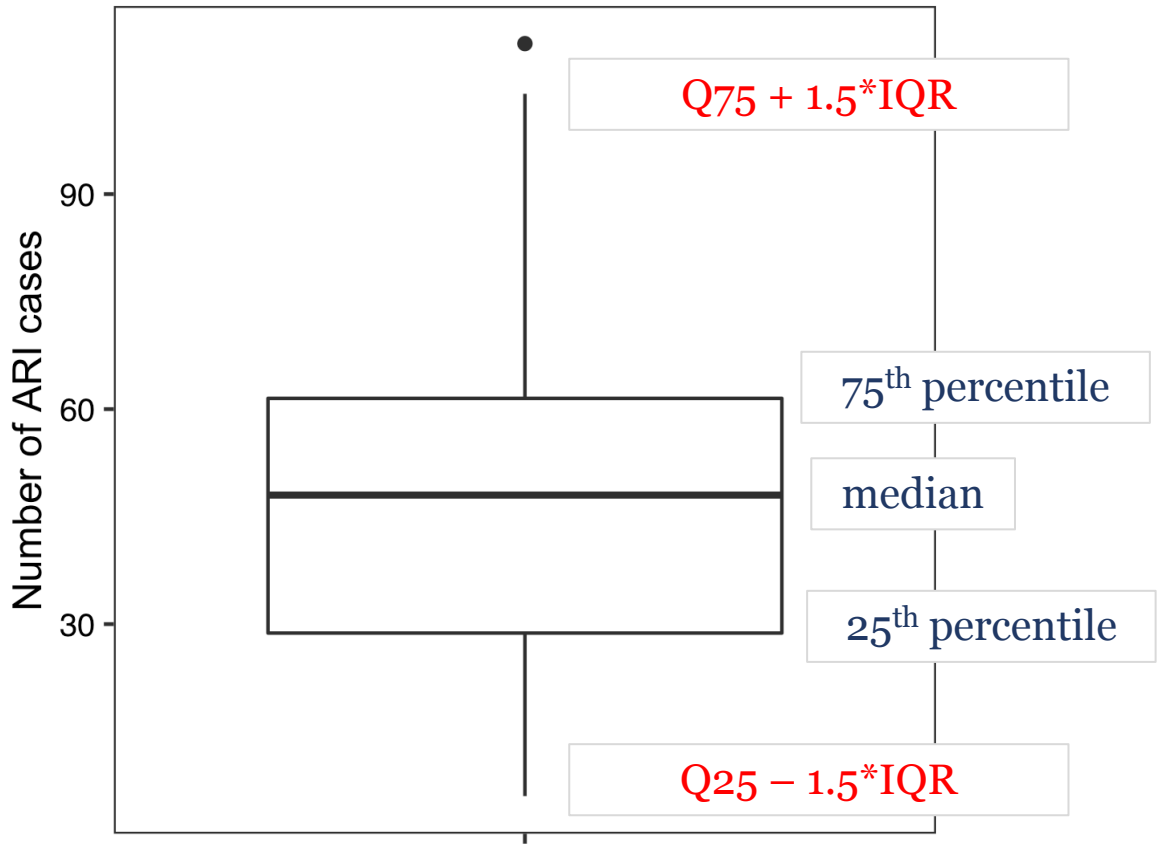




Boniken Clinic in Liberia

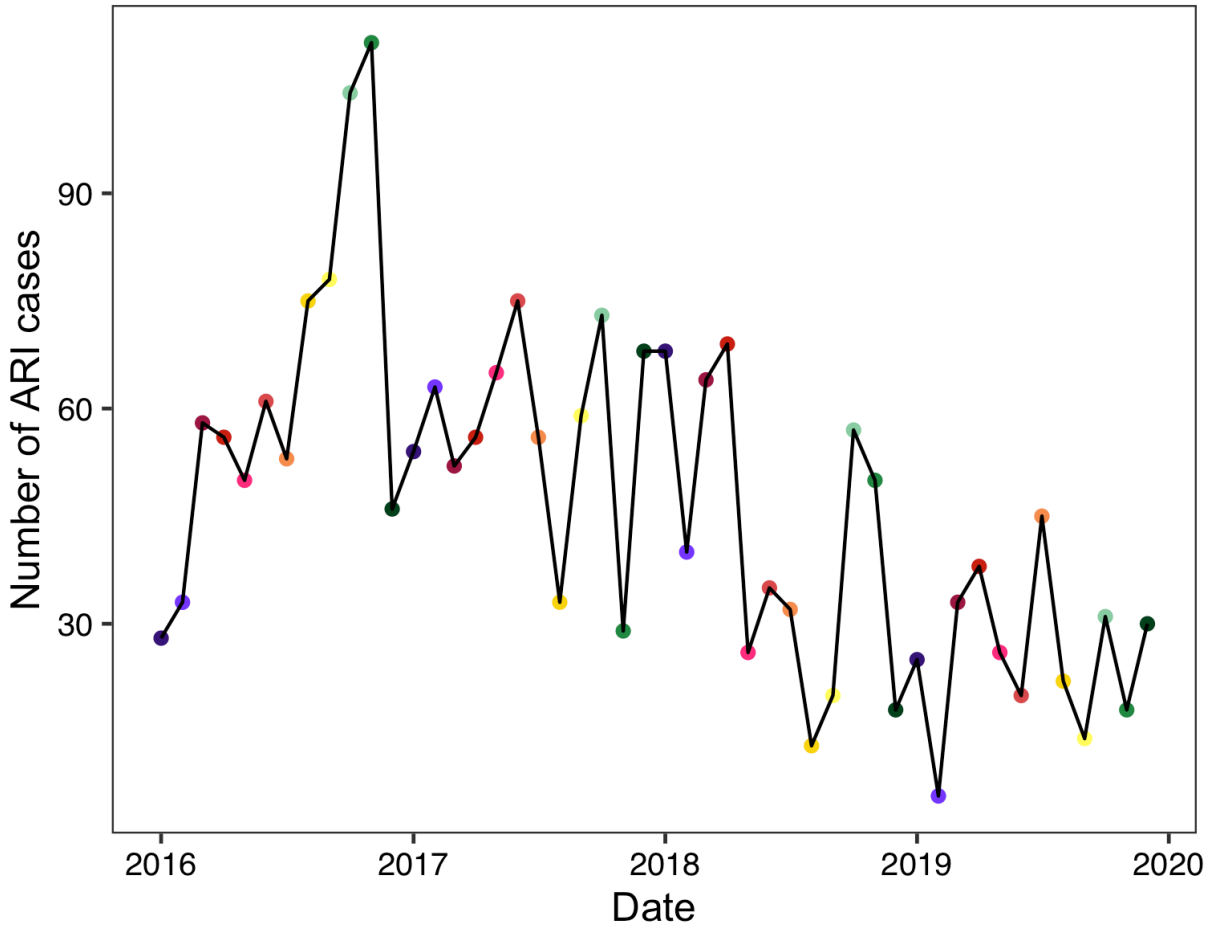
Method 2: Tukey's rule

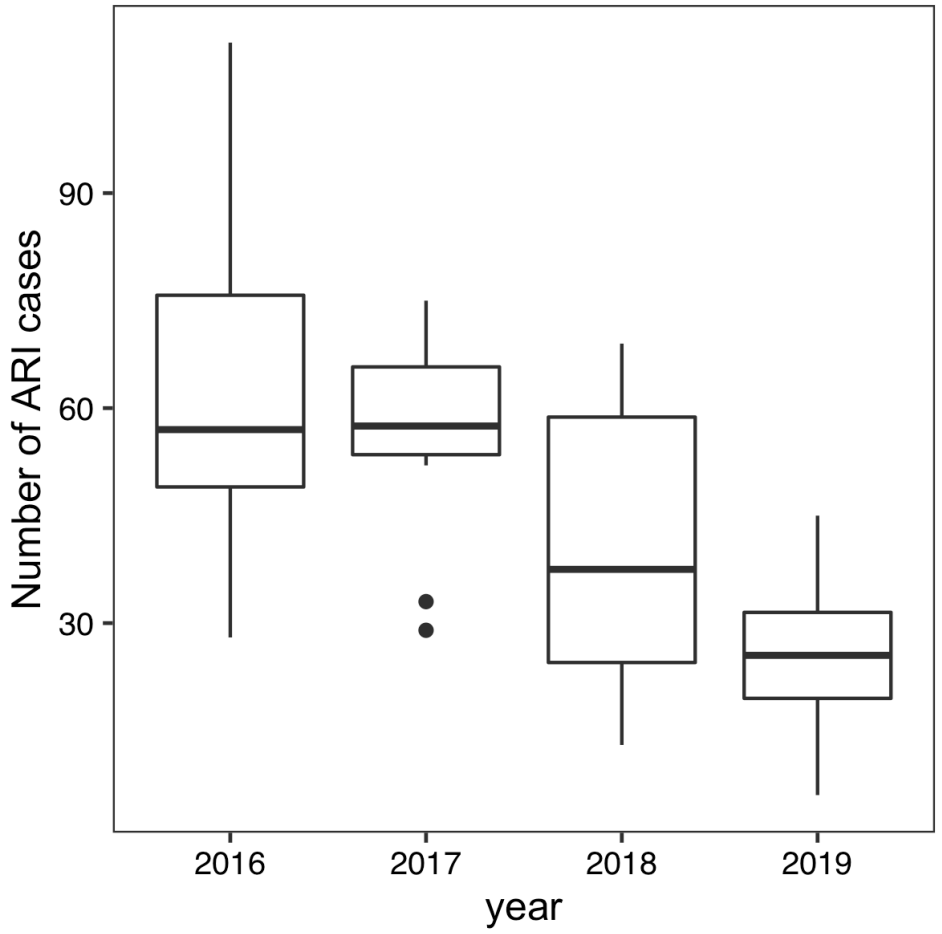




Boniken Clinic in Liberia

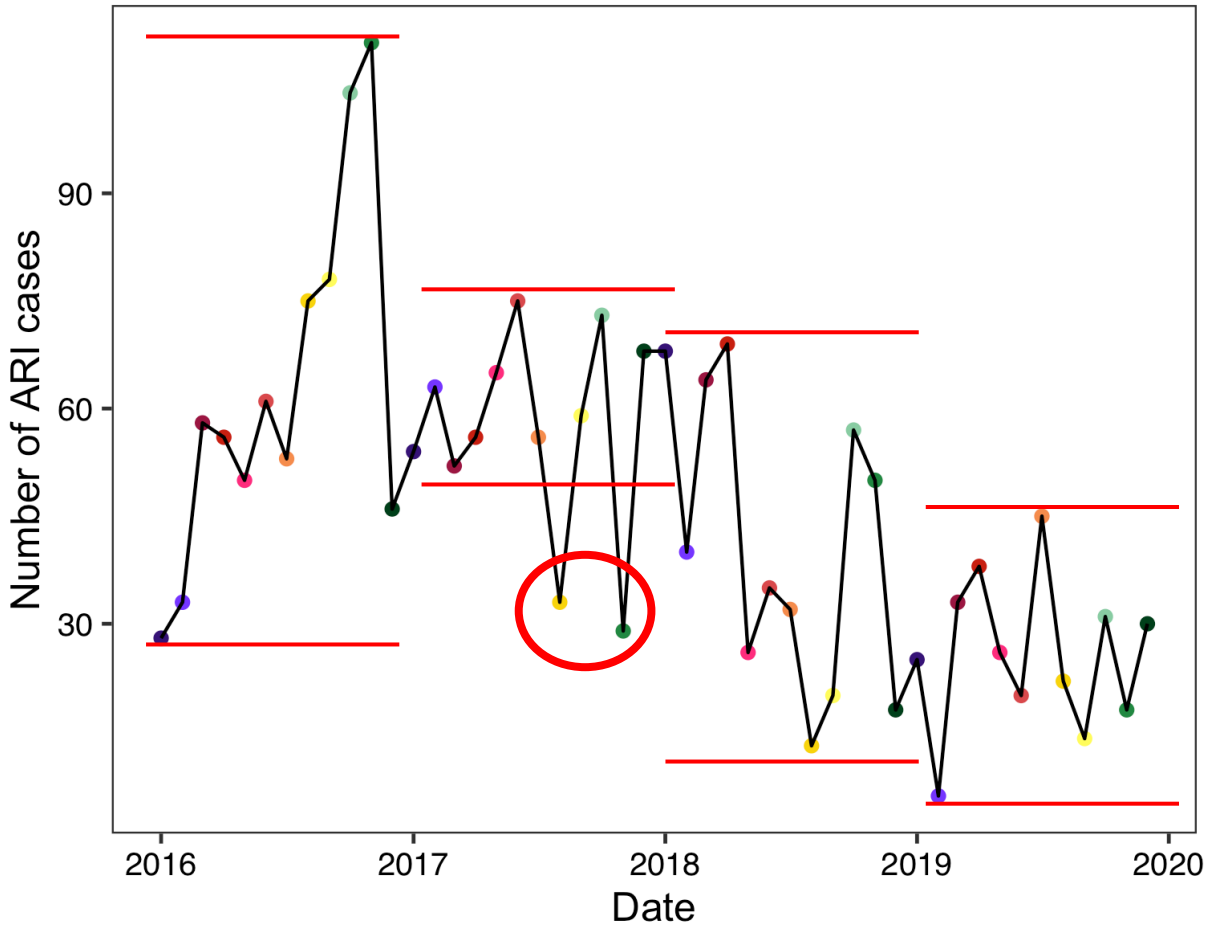
Method 2: Tukey's rule

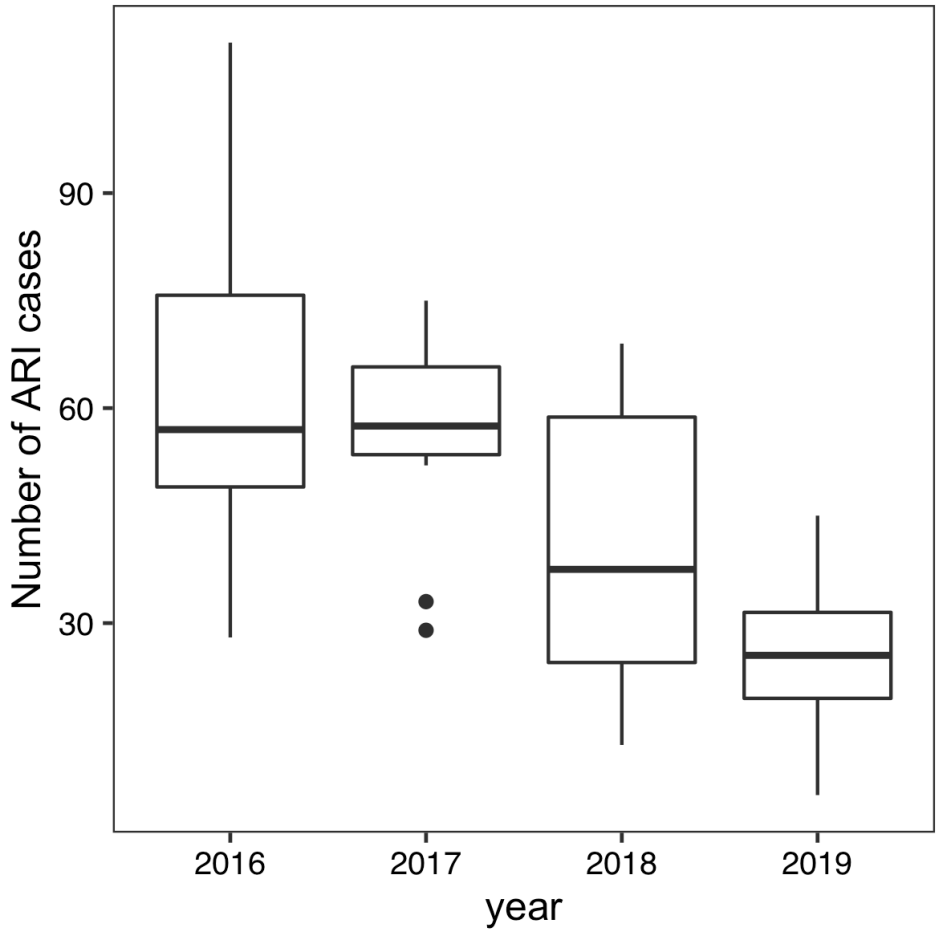




Boniken Clinic in Liberia

Method 2: Tukey's rule





Boniken Clinic in Liberia

Method 3: Apply statistical test

- **Fit time series model** *(Sessions 2 & 3)*
- **Calculate residuals** *(observed - predicted)*

$$r_t = y_t - \hat{y}_t$$

- **Calculate test statistic** *(Grubb's test)*

$$R_i = \max \frac{|r_t - \bar{r}|}{\sigma_r}$$

- **Calculate critical value**

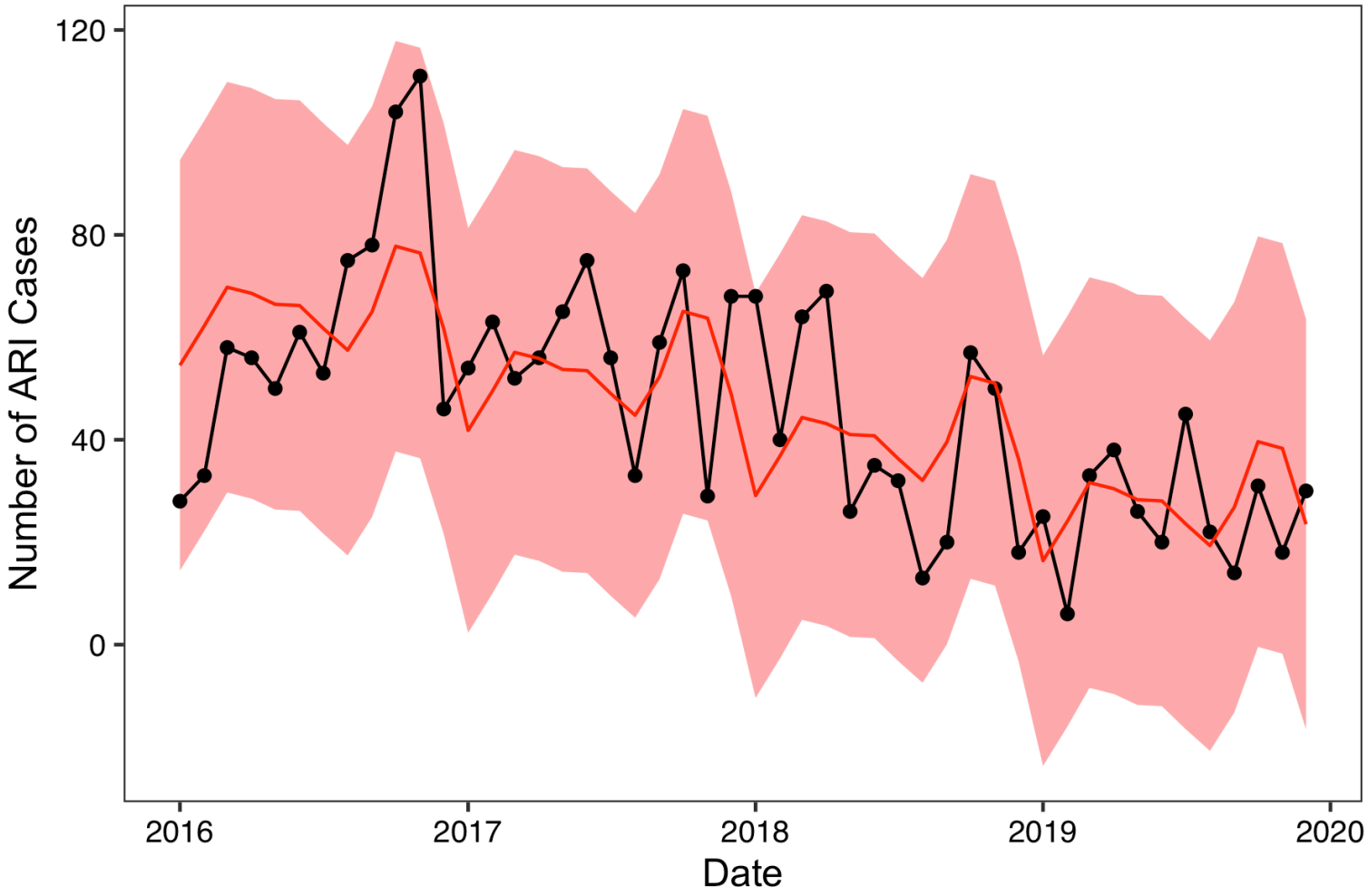
$$T_i = \frac{(N-i)t_{\alpha,T-i-1}}{\sqrt{(N-i-1+t^2_{\alpha,T-i-1})(N-i+1)}}$$

- **Compare test statistic to critical value**
  - *If larger, then remove the largest residual and repeat steps*
  - *If smaller, then no outliers identified*

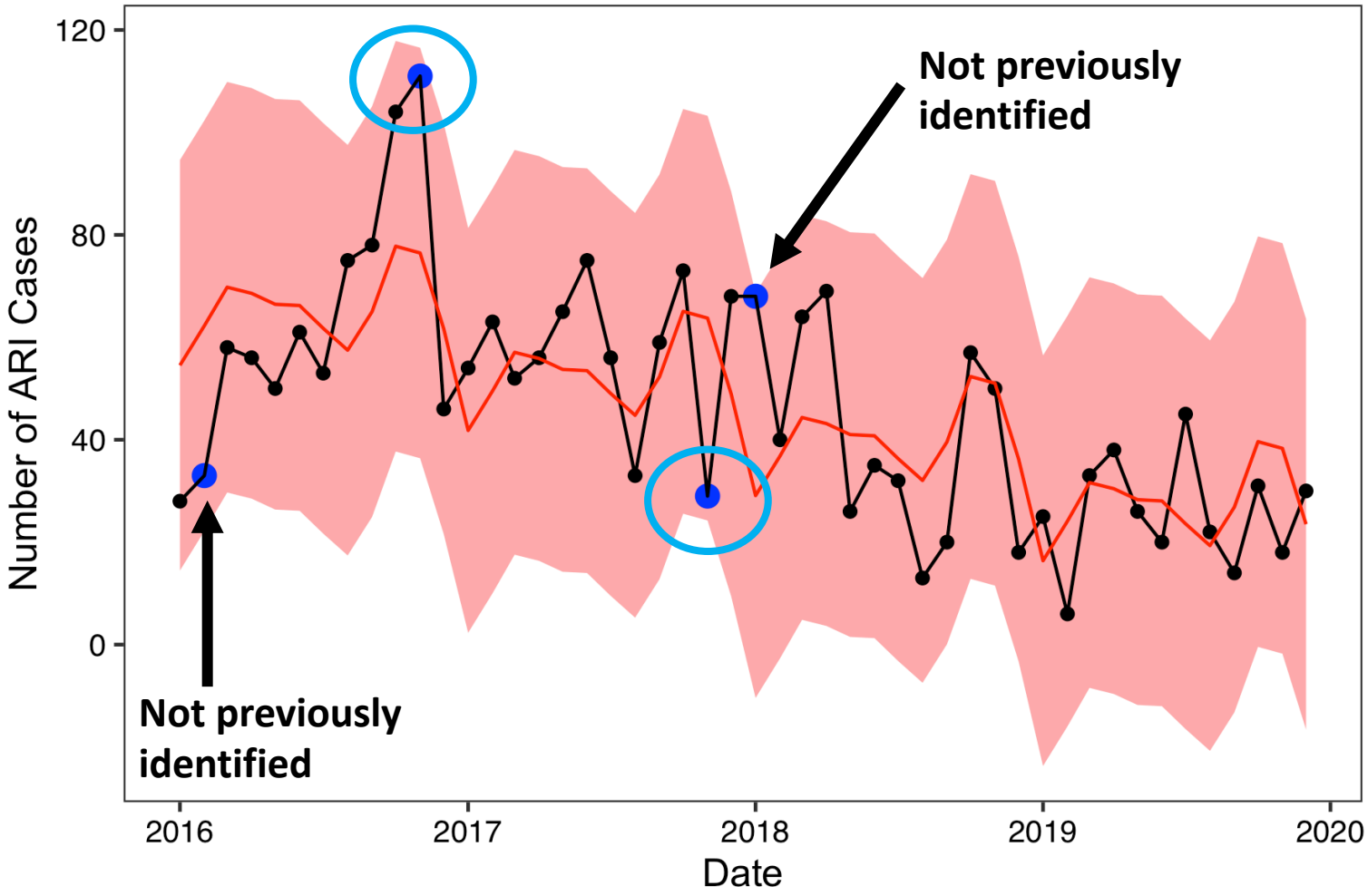
$$R_i > T_i$$

- **The residuals removed are the outliers**

Method 3: Apply statistical test



Method 3: Apply statistical test



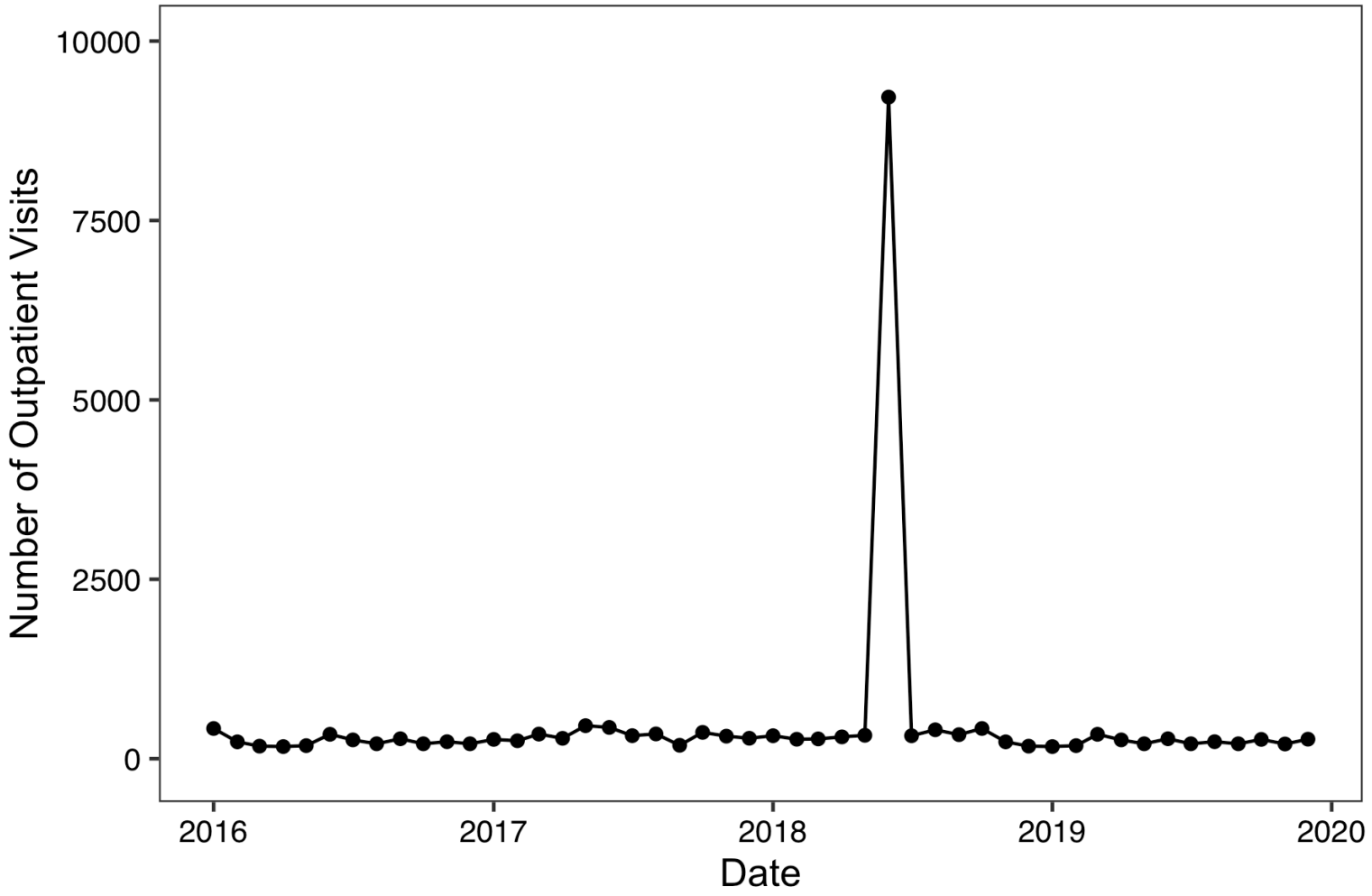
Solutions: Outliers

- **If an outlier is suspected, it is important to investigate *why* the data point may be an outlier**

- **Was it an anomaly?**
  - If concerned about model fit, include a dummy variable for the time point(s)

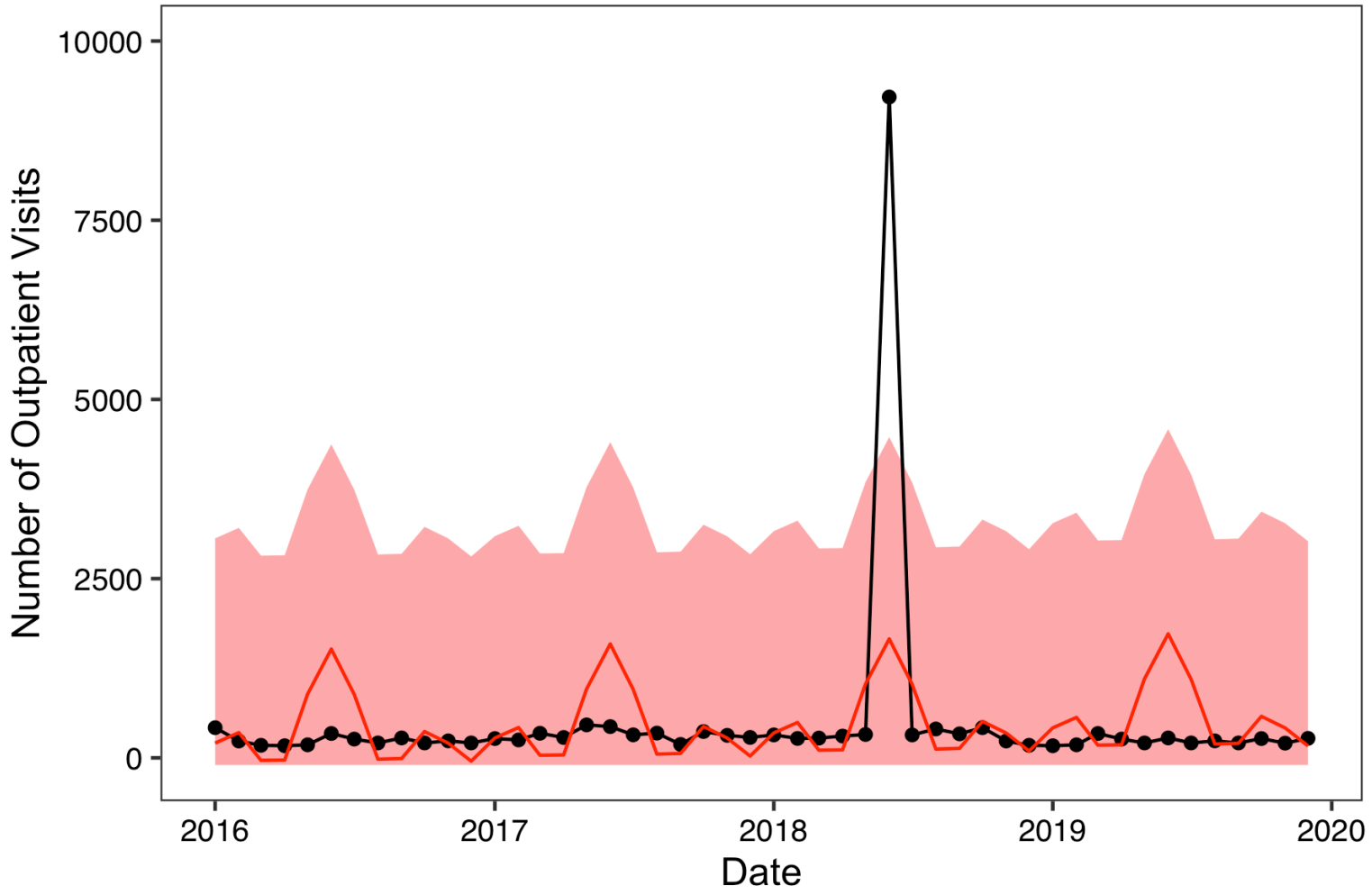
- **Was it a data entry error?**
  - Can it be replaced with the correct/true value?
  - If not, then code as missing value

Solutions: Data entry error



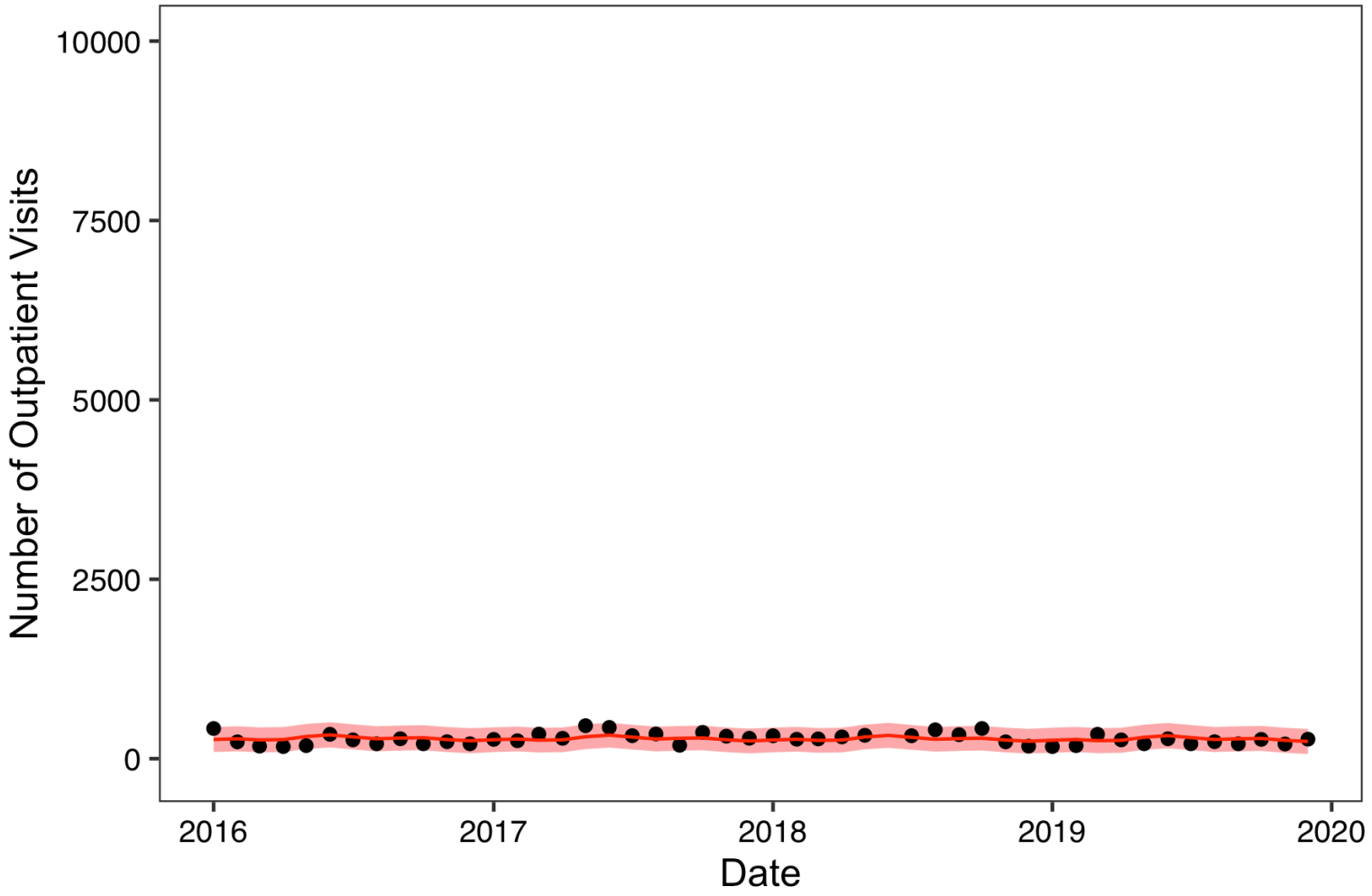
Jayproken Clinic in Liberia

Solutions: Data entry error



Jayproken Clinic in Liberia

Solutions: Data entry error



Jayproken Clinic in Liberia

Missing data

Considerations for missing data

To help determine if and how missing data should be addressed:

1. **What is the research question?**

2. **Why is the data missing?**

3. **Is there enough information to address missing data?**
   - How much data is missing?
   - Is there additional information available to “fill in” missing values?

1. What is the research question?

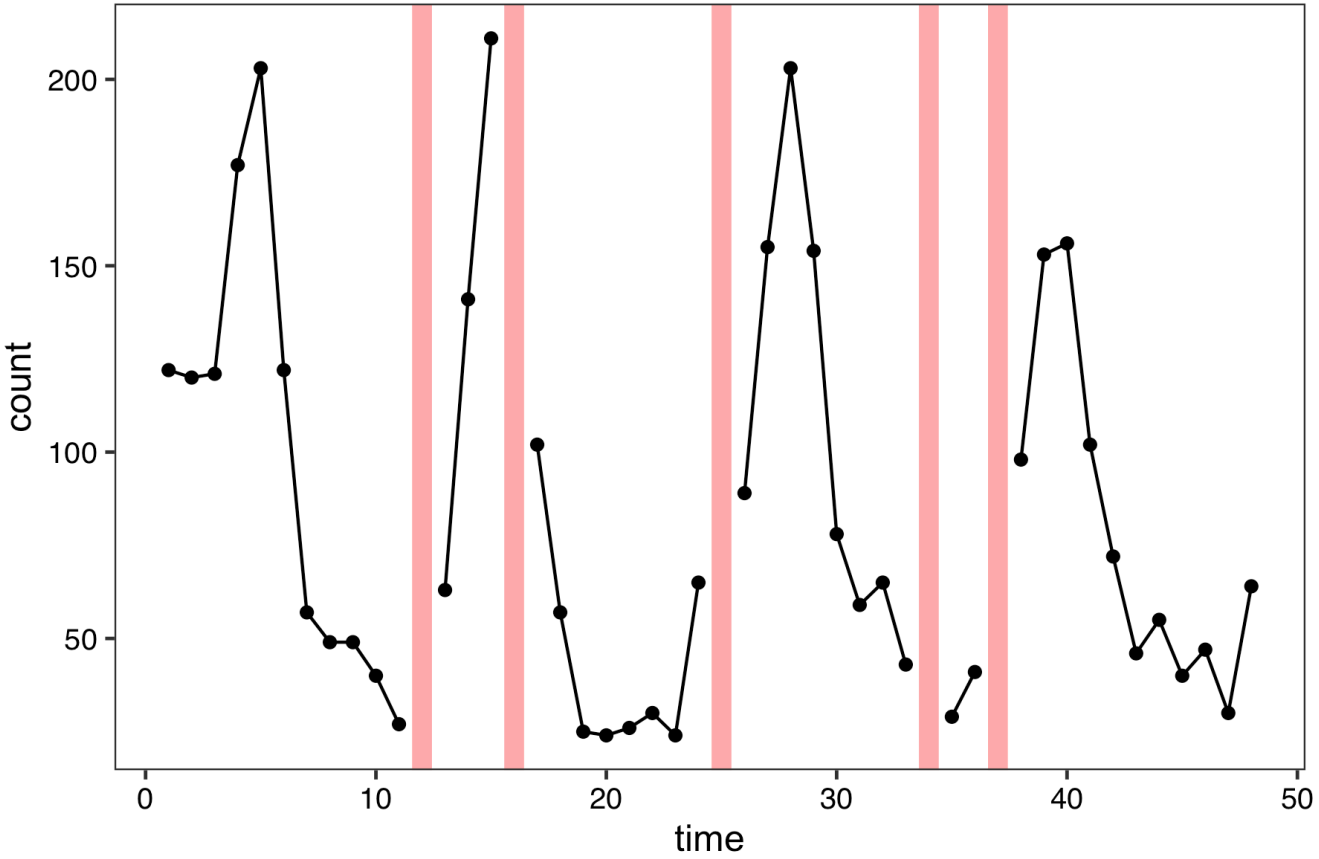
- **Describe** the behavior of an indicator over time
  - May be of interest to describe missing data pattern
  - May want to “fill in” a reasonable value for missing month in figures

1. What is the research question?

- **Describe** the behavior of an indicator over time

Can perform single imputation based on:

- Mean
- Interpolation
- Model-based

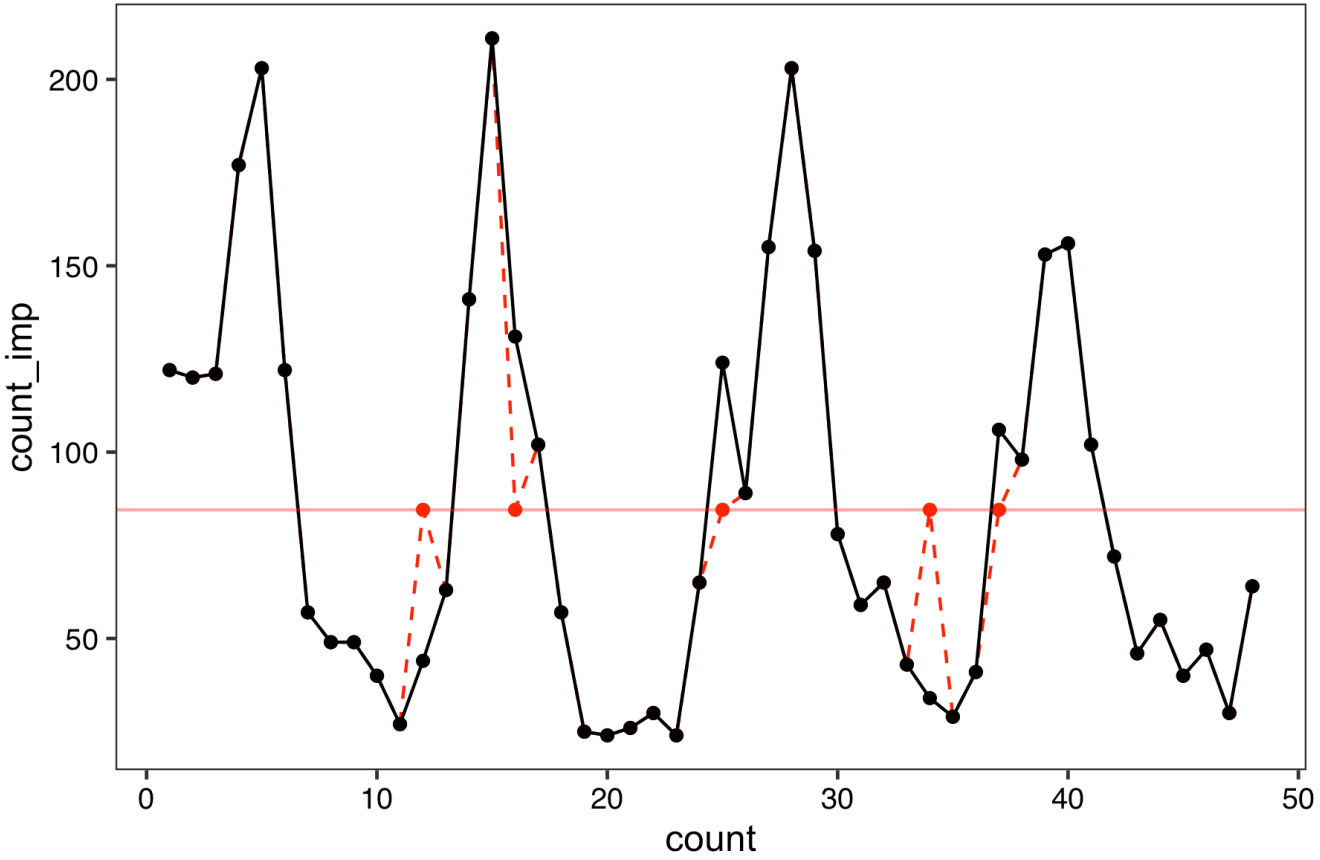


1. What is the research question?

- **Describe** the behavior of an indicator over time

Can perform single imputation based on:

- **Mean**
- Interpolation
- Model-based

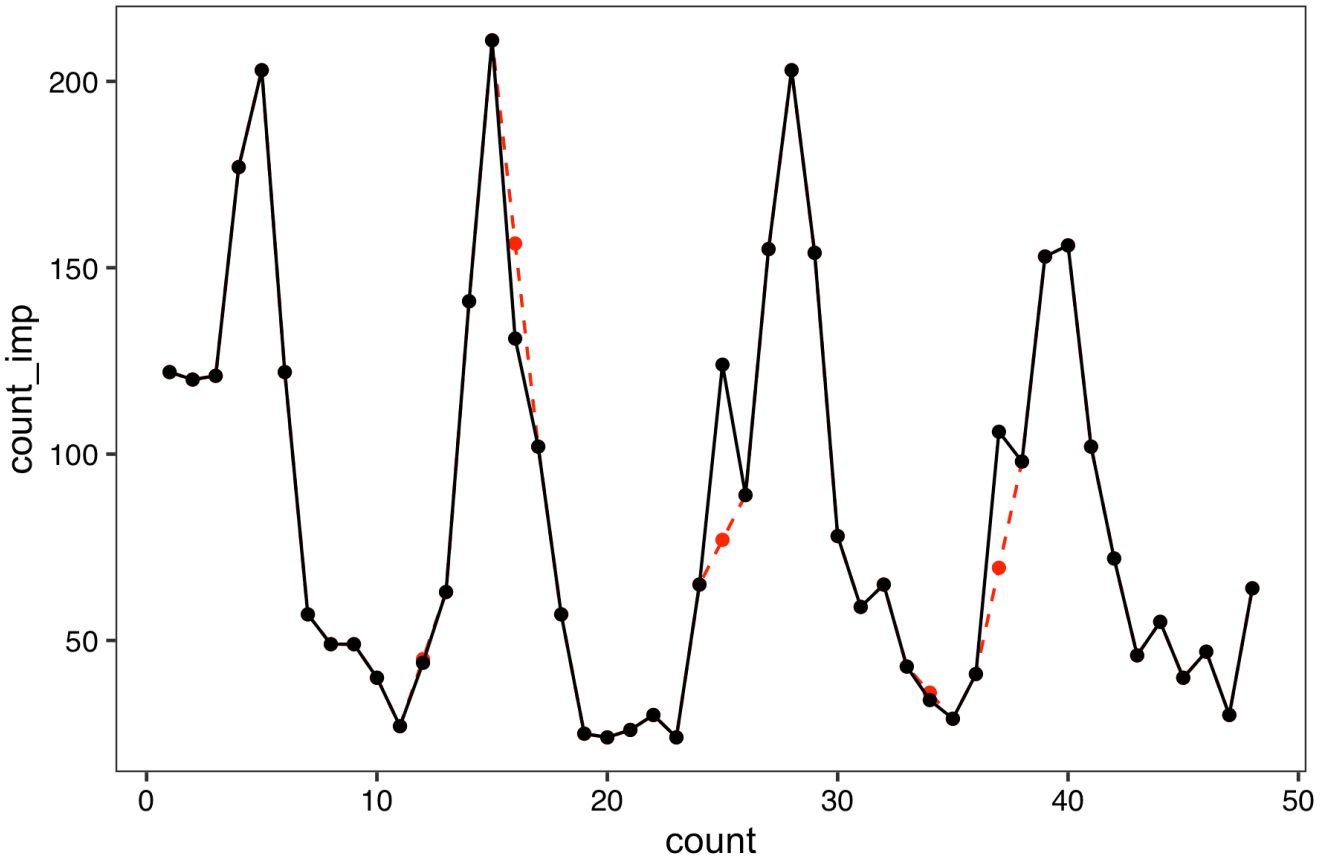


1. What is the research question?

- **Describe** the behavior of an indicator over time

Can perform single imputation based on:

- Mean
- **Interpolation**
- Model-based

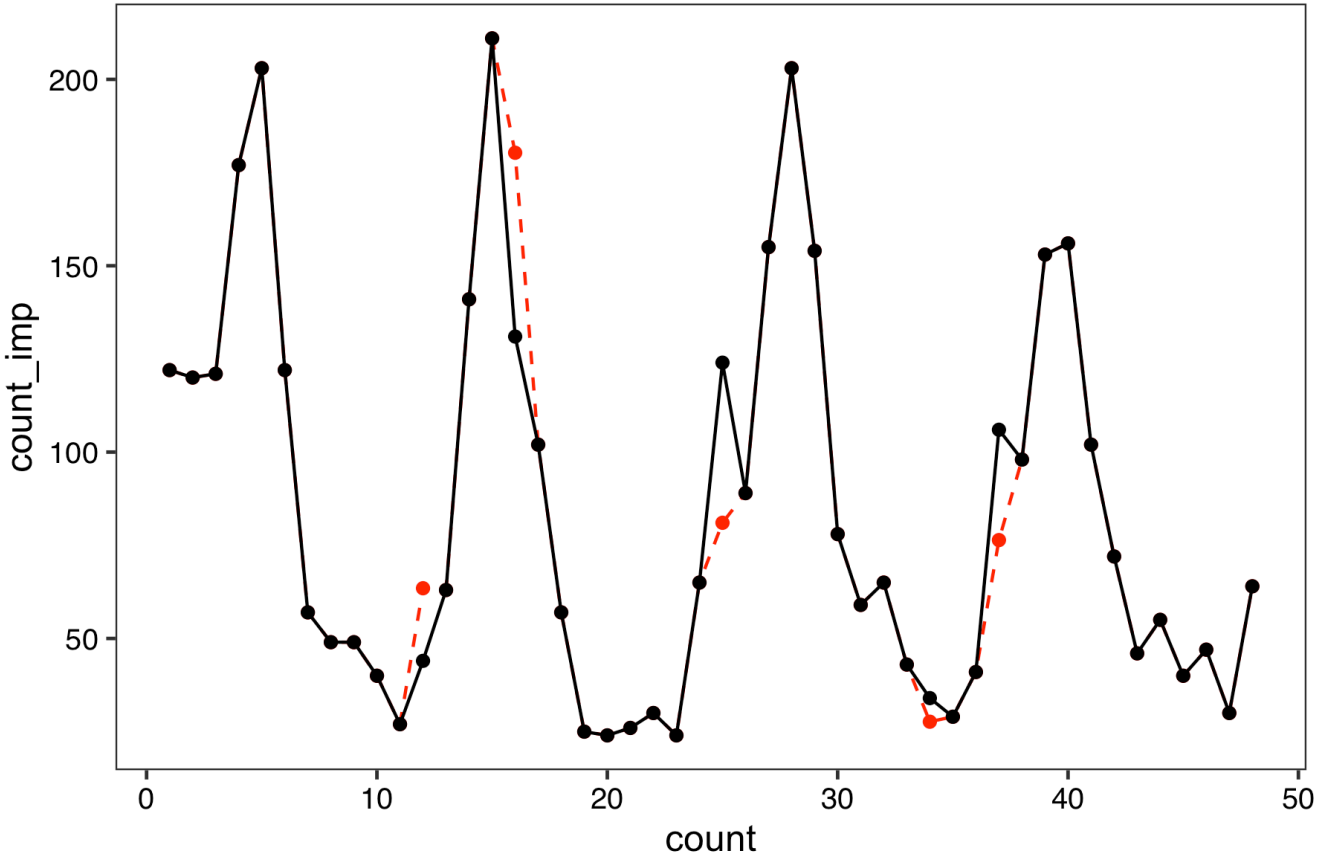


1. What is the research question?

- **Describe** the behavior of an indicator over time

Can perform single imputation based on:

- Mean
- Interpolation
- **Model-based**



1. What is the research question?

- **Describe** the behavior of an indicator over time
  - May be of interest to describe missing data pattern
  - May want to "fill in" a reasonable value for missing month in figures

- **Detect** deviations from expected in an indicator (*syndromic surveillance)*
  - Want to build a valid baseline model for each facility, which can be biased if data is missing
  - Want to aggregate data across multiple facilities
  - *Need to be careful about "borrowing" information from the baseline period for missing values in evaluation period*

- **Measure** the impact of an intervention on an indicator
  - Want to construct valid estimates, which can be biased if data is missing

2. Why is the data missing?

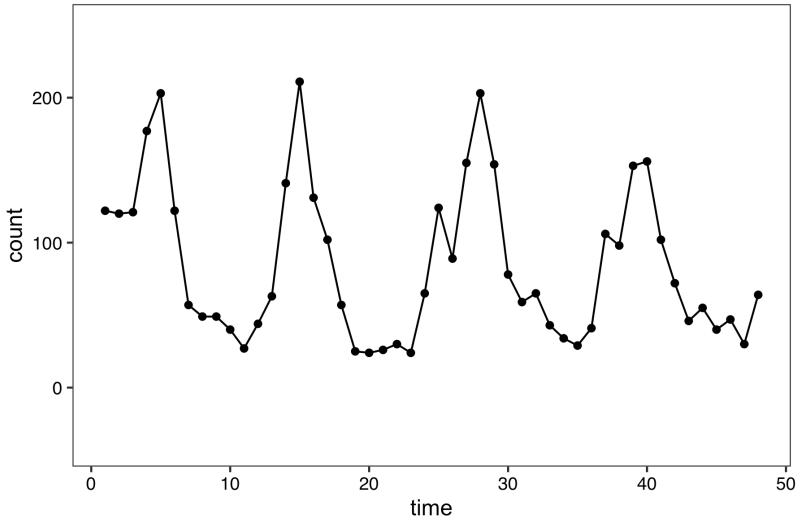
- **Missing completely at random**
  - Missing pattern is random (*can ignore: "complete case"*)

- **Missing at random**
  - Missing pattern can be fully identified using observed data (e.g. other facilities or external data sources)

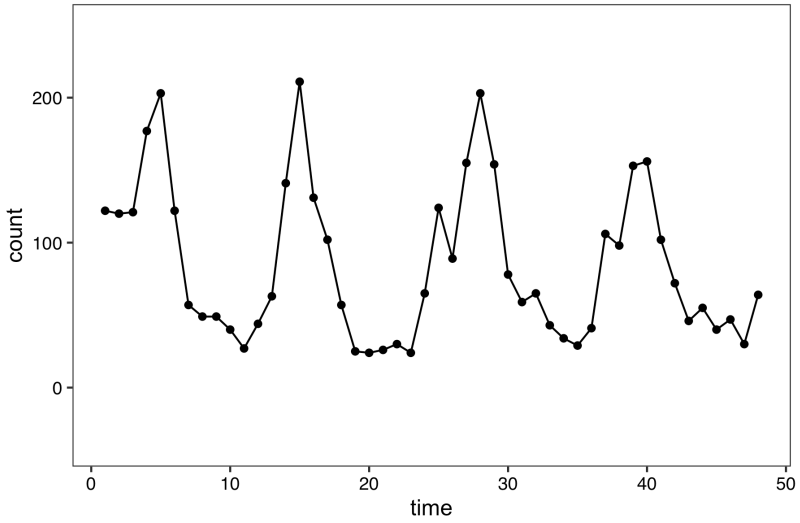
- **Missing not at random**
  - Missing pattern cannot be identified from the observed data
  - Very challenging and often results in calculating reasonable "bounds" for predicted values

Missing data patterns in time series data

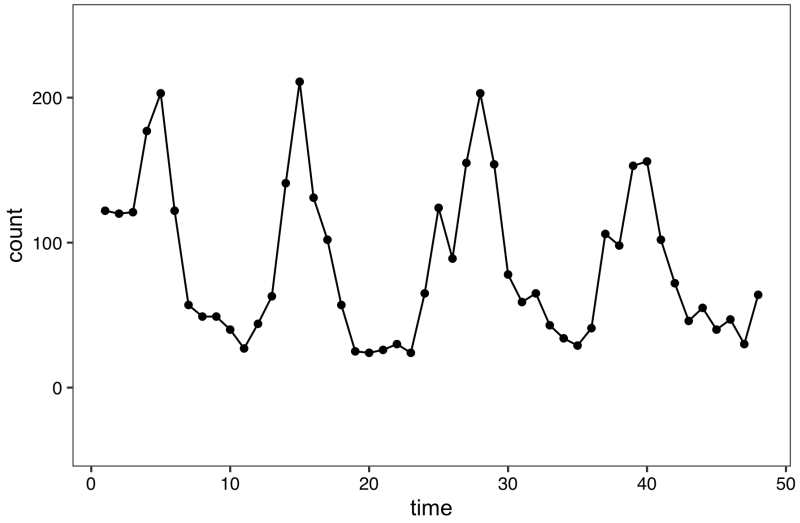
No missing data



MCAR

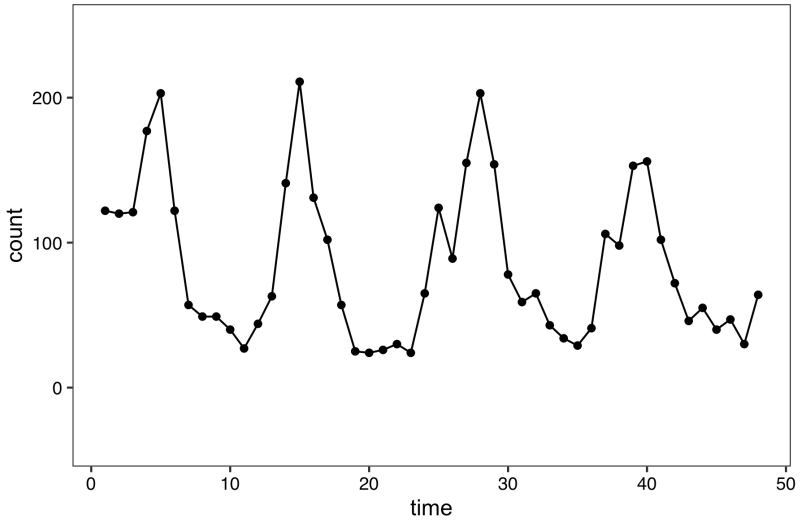


Not MCAR

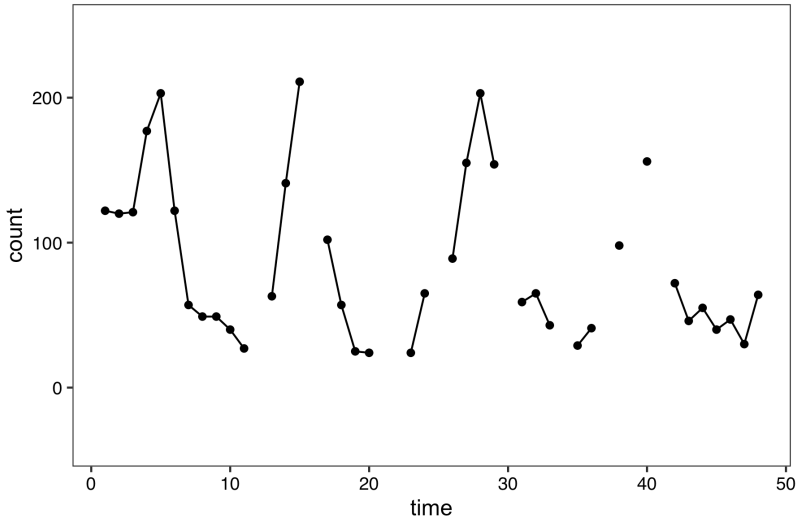


Missing data patterns in time series data

No missing data

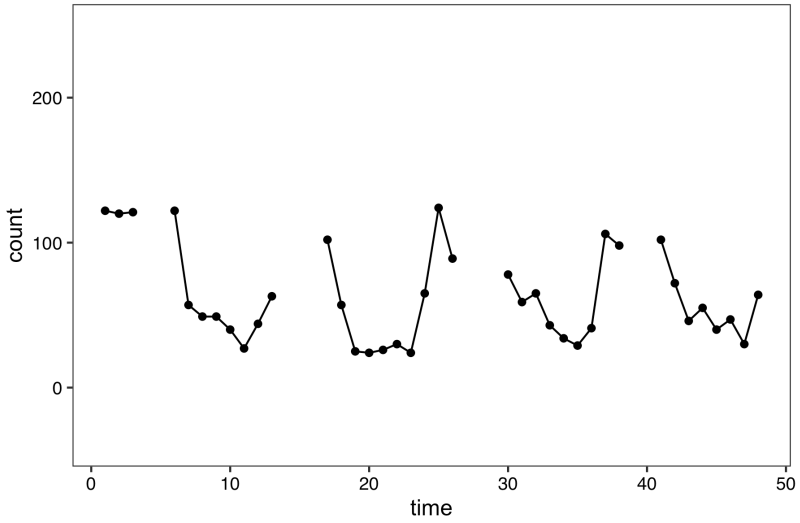


MCAR



remove 10 <u>random</u> points

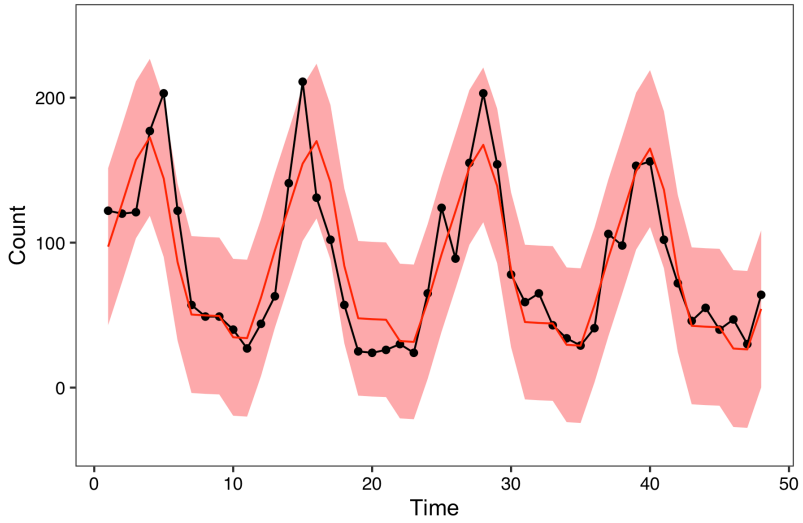
<u>Not</u> MCAR



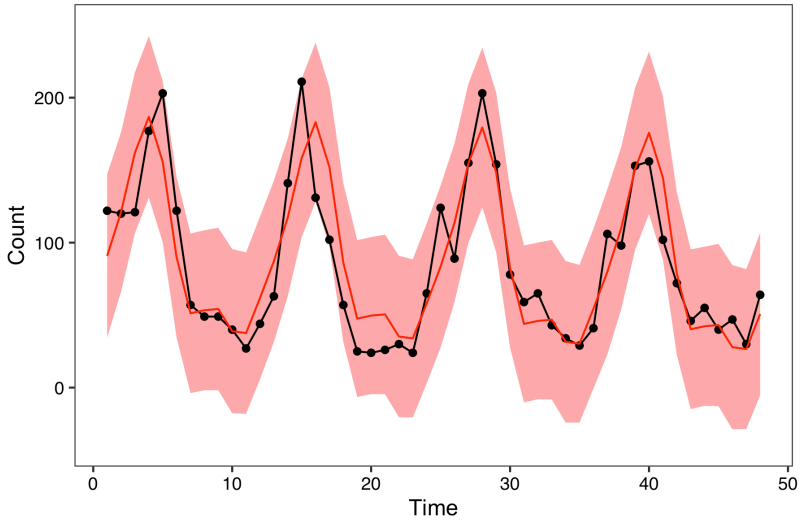
remove the 10 <u>largest</u> points

Missing data patterns in time series data

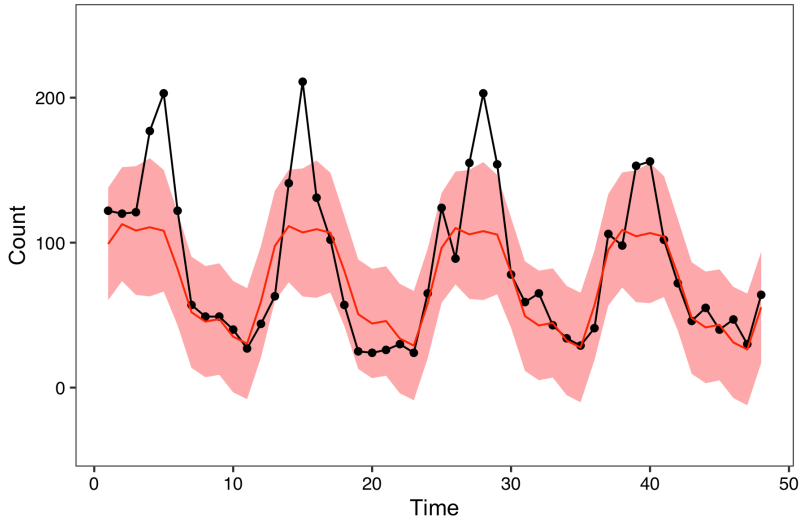
No missing data



MCAR

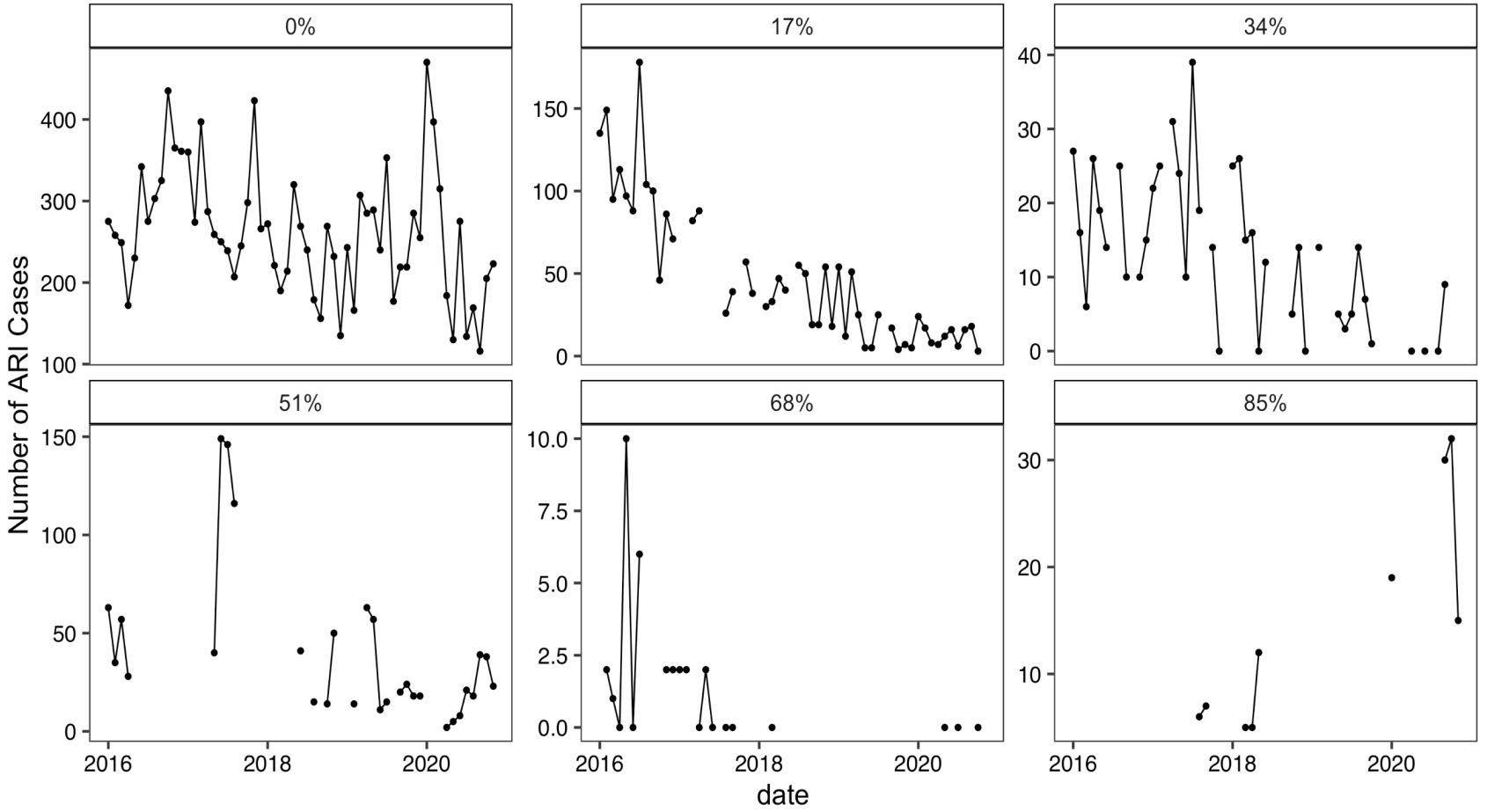


<u>Not</u> MCAR



3. Is there enough information?

- **How much data is missing?**



3. Is there enough information?

- **Is there additional information available to "fill in" missing values?**

- **Data from other facilities**
  - Can use "similar" facilities (neighboring, same size or type, or correlated seasonal patterns) to impute missing values

- **Other indicators**
  - Health service utilization indicators that are correlated with indicator could be used to impute missing values in a given facility
  - Need to make sure they do not have similar missingness patterns

- **External (non-DHIS2) sources**
  - Weather of rainfall data
  - Mobility data

Many statistical methods to incorporate these data sources!

Solution: Syndromic Surveillance

Goal 1: Create baseline facility-level models for prediction

- Chose not to model facilities with >20% of months missing
- This was "not enough information to draw conclusions"
- Fit facility-level models on complete case data

Goal 2: Create baseline aggregate-level models for prediction

- To calculate aggregate-level monthly predictions, sum predicted monthly counts across facilities by "drawing" from facility-level complete case models

Future direction: Utilize information from "similar" facilities and indicators to fill in missing values (*previous slide*)

Lab: Data formatting and cleaning

- **Nichole** will lead the lab today
- Work on cleaning and formatting data for analysis
- Examples with outliers and missing data

Final lecture: Data visualization for syndromic surveillance