Proof of concept

Table of contents


```
The downloaded binary packages are in
    /var/folders/2x/_ynr2g956zlgd037jy_156bm0000gn/T//RtmpUz11Pn/downloaded_packages
```

Introduction

This paper sets out an illustrated approach to developing an end-to-end approach to indicator development, publication and interpretation to assess health status, help develop health policy, and evaluate performance for the Saudi Public Health Authority (SPHA).

SPHA have selected 4 indicators for this proof-of-concept. They are:

1. Percentage of bloodstream infection due to methicillin-resistant Staphylococcus aureus (MRSA)
2. Percentage of bloodstream infection due to Escherichia coli resistant to 3rd-generation cephalosporin
3. Current cigarette smoking among women aged 18-44 years

4. Non-fatal Hospitalizations for All Injuries

A generic approach to indicators

Indicator production pipelines

It can be helpful to think of the indicator development and production process as pipeline which takes raw data as a input, and "pipes" it through various processing stages to generate indicators which can be analysed and visualised. At each step of the process are "inspection hatches" - quality assurance points whicu ensure that the process is doing what is expecged. Years of experience suggests a generic approach to developing actionable indicator sets. This is summarised in Figure 1.

Generic indicator definition

Metric a set of data or measures

Indicator A summary measure that aims to describe, in a few numbers as much detail as possible about a system, to help understand, compare, predict, improve, and innovate. (Pencheon 2007)

A key step it so recognise that indicators are numerical values calculated from a numerator and denominator and in a public health context, numerators are event or instance counts and denominators are populations at risk. Public health metrics are generally rates, ratios, or proportions.

$$Metric = numerator/denominator$$

Both numerators and denominators will be derived and aggregated from an underlying record system (for example clinical systems, lab systems, vital registration systems, survey databases) which may hold individual records, already aggregated data, or survey results.

The process of generating indicator values will require extracting the right values (business logic, e.g. SQL queries) as part of an Extract-Transform-Load (ETL). Ideally, indicator values would be calculated directly from numerator and denominator values (although this may not always be possible for example if metrics are age-standardised), and should flow and be stored at the **most disaggregated level possible without compromising confidentiality**. This ensures maximum flexibility and responsiveness to policy change, increases statistical power, enables trend analysis and underpins quality assurance.

Therefore a public health indicator system will need a regular flow of highly disaggregated numerator and denominator data from primary data sources. Efficiency dictates that these flows are standarised, and utility that they are as near-real-time as possible.

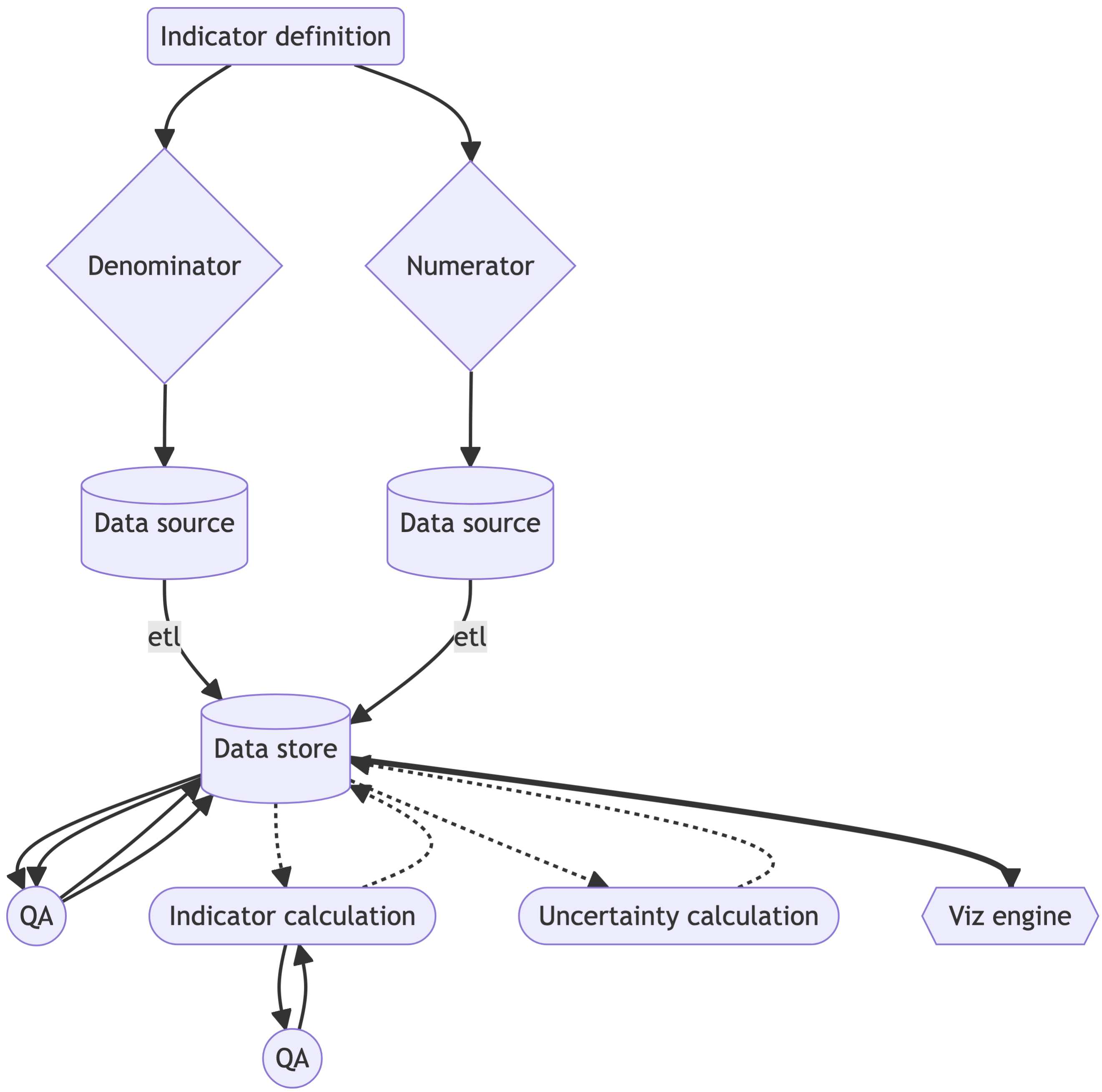


Figure 1: Simplified indicator production workflow

Good meta-data is essential for operating an indicator system and should be linked to each indicator and stored in database format. Each element needs a description of its source, how the value is derived (business logic or calculation details), method of calculation of uncertainty/confidence intervals, as well as rationale, stratifications e.g. temporal disaggregation, age-sex, SES breakdowns and additional information essential for interpretation for example, data discontinuities, recency and so on.

Illustration #1

E.coli cephalosporin resistance

To illustrate I will use English data for E.coli resistance to 3-rd generation cephalosporins. These data are an adjunct to the Public Health Outcome Framework and are publicly available for each NHS hospital in England on a quarterly basis. Detailed metadata is shown in Table 1.

Table 1: Detailed metadata

IndicatorID	definition	description
93188	Indicator	Rolling quarterly average proportion of 3rd generation cephalosporin resistant E. coli blood specimens; by quarter

IndicatorID	definition	description
		Percentage ofE. coliblood specimens (from English laboratories) with resistant test results fora3rd generation cephalosporin (in the first patient specimen in specimens successfully linked to mandatory surveillanceE. colibacteraemia reports);by acute Trust and by Quarter (2016 Q1 onwards). Presented as a rolling quarterly average across a 4 quarter period.The first quarter presented is Q4 2016, the quarterly average across 2016.3rd generation cephalosporin-resistance inE. colii

IndicatorID	definition	description
		This indicator demonstrates the proportion of patient linked mandatory and laboratory surveillance E. colibacteraemia episodes, presented by reporting acute Trust (mandatory surveillance)that have been tested for susceptibility to3rd generation cephalosporin each NHS acute Trust by quarter, in line with national guidance, and found resistant. A rolling quarterly average (across a 4 quarter period) is presented to comply with data protection regulations when dealing-

IndicatorID	definition	description
93188	Data source	Routine voluntary laboratory surveillance reports to UK Health Security Agency via the second generation surveillance system (SGSS); Antimicrobial testing data module
93188	Indicator source	

IndicatorID	definition	description
93188	Methodology	Thequarterly aver- ageper- centage of resistanta- ntibi- otictest results in patient linkedE. coli positive blood specimens recorded by an acute Trust setting in England. Total number of3rd generation cephalosporin resistant test resultson linkedE. colipositive blood specimens for the most recent 4 quarters is aggre- gated and divided by 4; creating the average quarterly value across the last year. This value is divided by the average quarterly value for

IndicatorID	definition	description
93188	Standard population/values	
93188	Confidence interval details	No confidence intervals generated for this indicator

IndicatorID	definition	description
93188	Source of numerator	Voluntary laboratory reports ofEscherichia coli(E. coli) blood specimens made by English laboratories to the routine laboratory surveillance scheme and extracted from UKHSA's Second Generation Surveillance System (SGSS) antimicrobial testing module which have been successfully linked (by patient) to a mandatory surveillance report of E. coli bacteraemia.

IndicatorID	definition	description
		The quarterly avera-genum-ber(across the last four quarters) of laboratory reports ofEscherichia coli(E. coli) from mandatory surveil-lance linked blood specimens submitted by English laborato-ries, tested for and are resistant to a 3rd generation cephalosporin.Data sum-marises the rolling quarterly average number of reports of3rd generation cephalosporin resis-tanttest results-from mandatory surveil-lance linkedEscherichia coli(E. coli)blood specimens submitted

IndicatorID	definition	description
93188	Source of denominator	Voluntary laboratory reports ofEscherichia coli(E. coli) blood specimens made by English laboratories to the routine laboratory surveillance scheme and extracted from UKHSA’s Second Generation Surveillance System (SGSS) antimicrobial testing module which have been successfully linked (by patient) to a mandatory surveillance report ofE. colibacteraemia.

IndicatorID	definition	description
		Rolling quarterly average (for the last four quarters) number ofinitial patient laboratory reports ofEscherichia coli(E. coli) identified from blood specimens made by English laboratories to the routine (voluntary) laboratory surveillance scheme and extracted from UKHSA's Second Generation Surveillance System (SGSS) that have been successfully linked to a mandatory surveillance E. coli bacteraemia report, and tested for susceptibility to a 3rd

IndicatorID	definition	description
93188	Disclosure control	

IndicatorID	definition	description
		Data used for this indicator are from the routine voluntary laboratory surveillance scheme following successful linkage (by patient identifier) to the acute Trust level mandatory reports. Total numbers ofE. colibacteraemia presented in this indicator differ from those presented in the mandatory surveillance scheme indicators due to reduced ascertainment in the laboratory reporting scheme.Linkage methodology improvements have been made and used for

IndicatorID	definition	description
93188	Copyright	© Crown copyright
93188	Data re-use	
93188	Links	
93188	Indicator number	
93188	Notes	
93188	Frequency	Indicator is updated on a quarterly basis
93188	Rounding	
93188	Indicator Content	
93188	Specific rationale	
93188	Simple Name	
93188	Simple Definition	
93188	Impact of COVID-19	
93188	Unit	%
93188	Value type	Proportion
93188	Year type	Calendar
93188	Polarity	RAG - Low is good
93188	Date updated	14/03/2024

The metadata provides detail for transparency, and also complies with FAIR principles for reporting (Rothfritz 2019). High quality meta-data is essential for data sharing and reproducibility.

- *Polarity.* This refers to the interpretation and directionality of high (or low) indicator values, specifically if a whether significant value are considered desirable or undesirable i.e. *low is good* or *high is good.* This concept drives the use of colour in data visualisation. For example, low mortality overall, or from any given cause is more desirable than high mortality. Note that mortality is generally compared with a benchmark, and statistically significant difference determined by whether the confidence interval of the indicator contains the benchmark value. If an metric has a definite polarity, it has more use an actionable indicator than a metric without polarity. The PHOF makes extensive use of RAG (red-amber-green) colour palette to denote polarity.
- *Red-red.* If comparisons are made at different scales, of for both temporal and cross-sectional comparisons, a red-red approach is a rapid surveillance tool for detecting strata that are both performing badly, and not making progress. (See example)

Resistance data in England is published quarterly at hospital level and is calculated as rolling year i.e. each point estimate is the annual value to that point in time. This approach smooths the data.

Data is currently available from 2016 Q1 to 2023 Q4 starting with Q4 2016.

Trend and variation

Data for England suggests that resistance rates have increased as has the variation between hospitals.

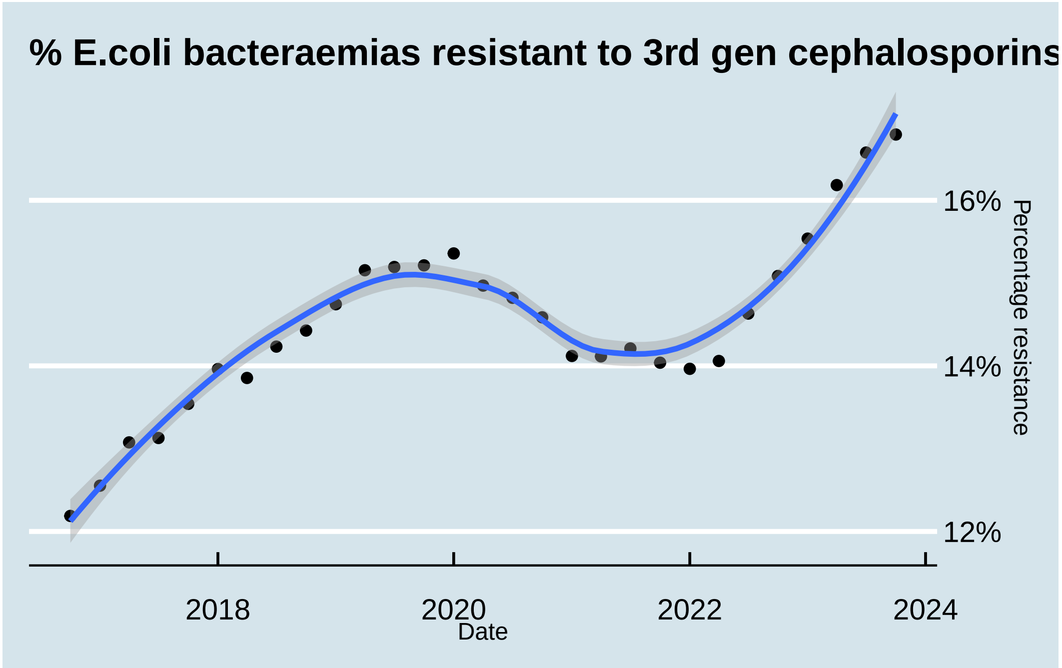


Figure 2: England Trend in % of E.coli bacteraemias resistant to 3rd generation cephalosporin

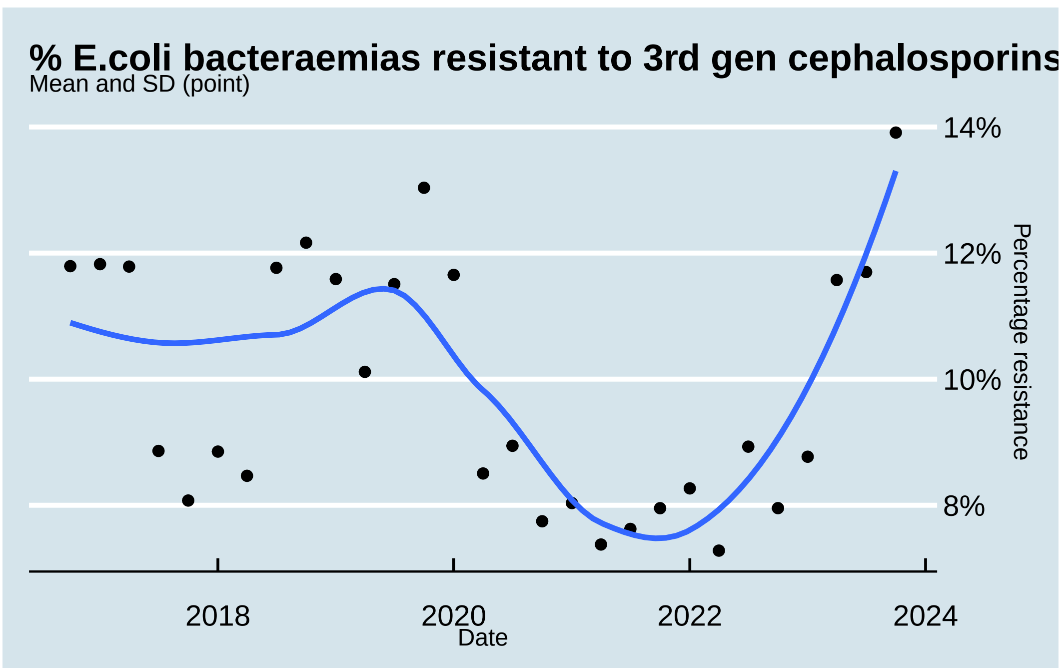
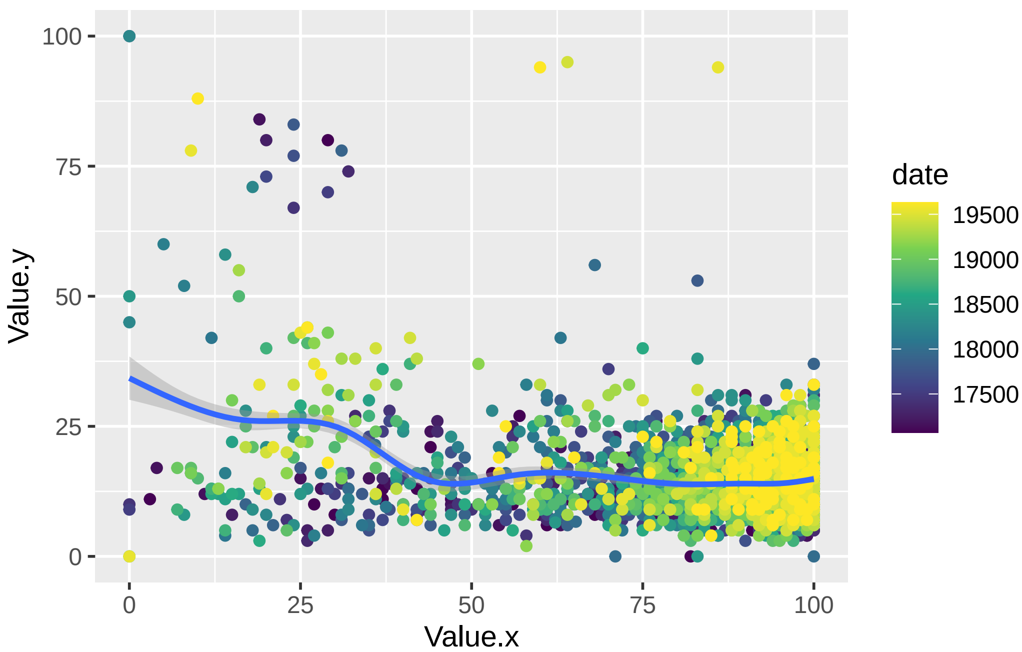


Figure 3: Mean and variation (SD)



All health indicators are measures of performance and performance is only useful through comparison. The level of comparison determines … For example to address the question "How is SA doing internationally" requires comparison with national metric values from comparator countries or regions. To address the question "are we improving?" requires two sets of time series data - the trend for SA for relevant metrics, and the trend for comparator areas or expected trend - which addresses the question "are we improving as expected / fast enough". The latter question is particularly relevant in relationship to inequalities where the aim might be to reduce the gap between populations.

Using dummy data for cephalosporin resistance

For this illustration we will simulate a dataset of 50000 bacteraemia observations for 8 regional sub-divisions within Saudi Arabia with variable AMR rates from 3 - 21%.

Table 2: First 20 rows of simulated results of E.coli baacteraemia resistance to 3rd gen cephalosporins

area	date	test	id
Area6	2019-07-01	No	1
Area2	2021-07-01	No	2
Area7	2023-10-01	No	3
Area3	2021-04-01	No	4
Area7	2019-10-01	Yes	5

Table 2: First 20 rows of simulated results of E.coli baacteraemia resistance to 3rd gen cephalosporins

area	date	test	id
Area8	2021-04-01	Yes	6
Area3	2022-07-01	No	7
Area4	2022-01-01	No	8
Area4	2021-04-01	No	9
Area6	2022-10-01	No	10
Area8	2019-07-01	No	11
Area6	2020-10-01	No	12
Area2	2020-07-01	No	13
Area2	2023-04-01	No	14
Area5	2020-01-01	No	15
Area2	2023-04-01	No	16
Area5	2019-04-01	No	17
Area7	2022-07-01	No	18
Area8	2022-04-01	No	19
Area1	2021-07-01	No	20

Using our modelled data we can quarterly trends in resistance and examine period to period change in variation (see Figure 4; Figure 5)

Illustration #2: Tobacco smoking

I have extracted smoking prevalence estimates for Saudi Arabia from the GBD Epi VIsualisation Tool. This contains prevalence estimates for smoking by sex and age for 1996, 2007, 2010, 2014. The main sources for these estimates are the Saudi Arabia Family Health Survey 1996-1997, Saudi Arabia Global Youth Tobacco Survey 2007 and 2010, WHO Report on the Global Tobacco Epidemic 2019.

The latest estimates from this source are now 10 years out of date but suggested that smoking prevalence in women was declining and was generally below 5% whereas in men, prevalence was generally increasing especially in younger age groups but with a wide variation between age groups (see Figure 6).

The more contemporary Saudi-PURE study (2012 - 2015) - a household survey based in Riyadh and central Saudi Arabia, found smoking prevalence in 35-70 year olds of 21% and 4% in men and women respectively (Alhabib et al. 2020).

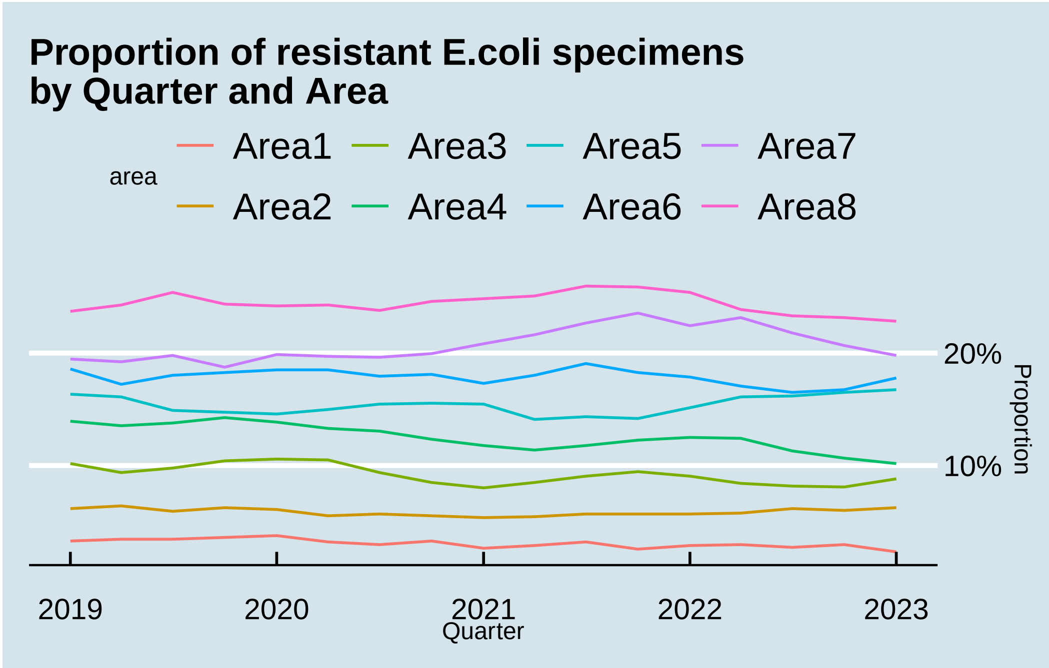


Figure 4: Trend

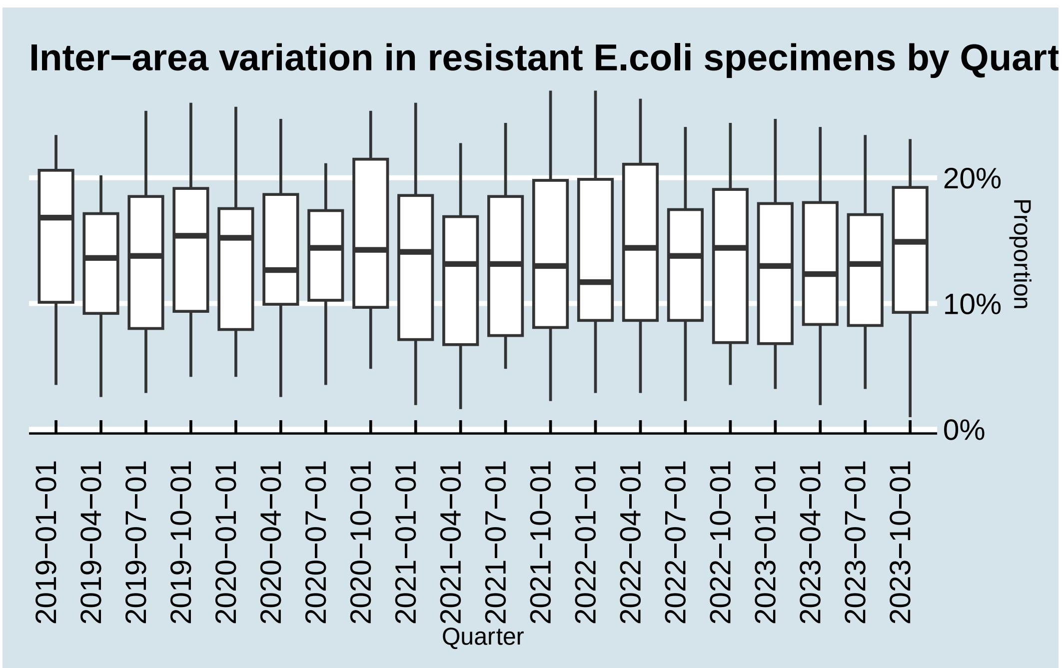


Figure 5: Variation

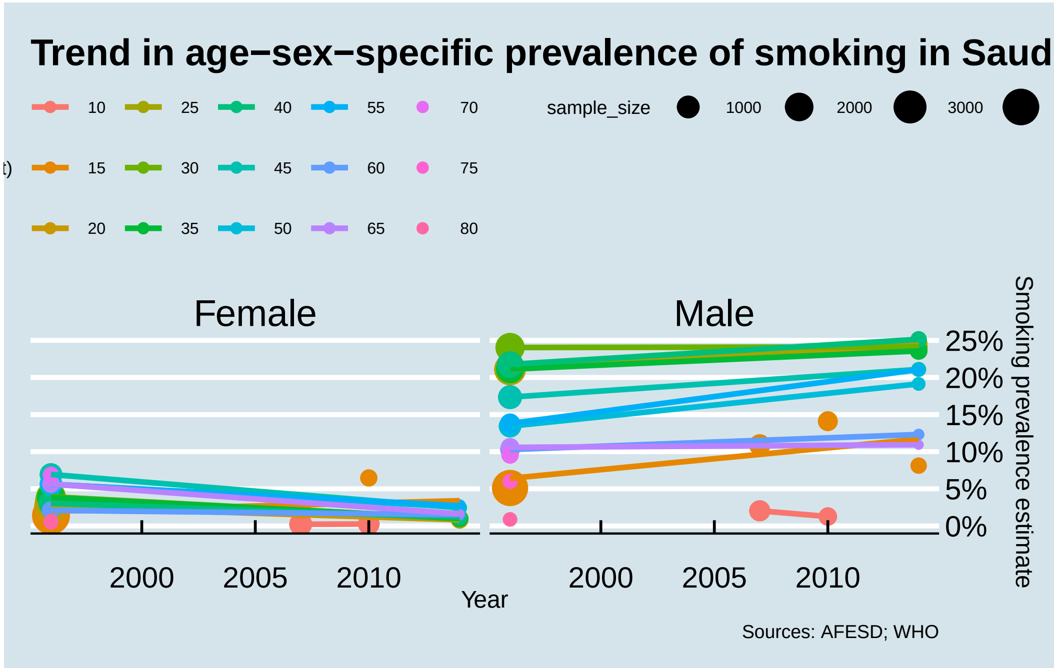


Figure 6: Trends in smoking by age and sex

Interpreting change

If we find a difference in an indicator value between two points in time, or between two population units, there are three potential explanations:

- The play of chance
- Data artefacts
- Genuine difference

The play of chance can be assessed by calculating uncertainty intervals for our metrics. Data artefacts can be much more difficult to detect and requires understanding of the indicator construction, examining trends in numerators and denominators, and variation for outliers and discontinuities. If differences are not due to chance or data issues, they are likely to reflect a real phenomenon. If differences are real, further investigation may require examining stratified data sets e.g. disaggregation by age or sex or area.

Therefore, to be able to disentangle or identify causality data should be as disaggregated in time, place and person as possible (without compromising confidentiality).

This may not always be possible - for example we may have annual surveys…

For some metrics may take many years to develop meaningful trends

Policy analysis

In an ideal world we would devise metrics which responded quickly to policy change in order to give rapid feedback on whether our interventions are working. We need to be able to understand the counter-factual - what would have happened without intervention.

Question	Comparison	Analysis
What is the problem?		
What is the variation?		
Are there outliers?		

Alhabib, Khalid F., Mohammed A. Batais, Turky H. Almigbal, Mostafa Q. Alshamiri, Hani Altaradi, Sumathy Rangarajan, and Salim Yusuf. 2020. "Demographic, Behavioral, and Cardiovascular Disease Risk Factors in the Saudi Population: Results from the Prospective Urban Rural Epidemiology Study (PURE-Saudi)." *BMC Public Health* 20 (1). https://doi.org/10.1186/s12889-020-09298-w.

Pencheon, D. 2007. "The Good Indicators Guide: Understanding How to Use and Choose Indicators." In. https://www.semanticscholar.org/paper/The-good-indicators-guide%3A-understanding-how-to-use-Pencheon/f65b05a6040f3ee456271a02987749d31d7c472c.

Rothfritz, Laura. 2019. "The FAIR Data Principles," March. https://doi.org/10.14293/s2199-1006.1.sor-compsci.clnbrup.v1.