How to evaluate healthcare policy: from data to interpretation

Imperial College London

How to evaluate healthcare policy: from data to interpretation

Midlands analyst network huddle Apr 2026

Alex Bottle

Professor of Medical Statistics
Imperial College London

Session overview

- Key statistical methods for evaluating any healthcare policy, from the roll-out or update of vaccines, new surgical technology, or the implementation of NICE guidelines: before v after comparisons, interrupted time series, and statistical process control charts
- How to frame the research question, prepare the data, choose the analysis, and interpret the output

Learning outcomes for the session

- Know the main stats methods for evaluating healthcare policy
- Understand their data requirements, strengths and weaknesses
- Be able to interpret their output and decide how successful the policy has been
- I’m not going to go into formulae or software

1: FRAMING THE ANALYTICAL QUESTION

Routes to diagnosis of heart failure: observational study using linked data in England

Alex Bottle [1], Dani Kim [1], Paul Aylin [1], Martin R Cowie [2], Azeem Majeed [1], Benedict Hayhoe [1]

Correspondence to Dr Alex Bottle, Department of Primary Care and Public Health, Imperial College, London, UK; robert.bottle@imperial.ac.uk

Abstract

Objective Timely diagnosis and management of heart failure (HF) is critical, but identification of patients with suspected HF can be challenging, especially in primary care. We describe the journey of people with HF in primary care from presentation through to diagnosis and initial management.

Methods We used the Clinical Practice Research Datalink (primary care consultations linked to hospital admissions data and national death registrations for patients registered with participating primary care practices in England) to describe investigation and referral pathways followed by patients from first presentation with relevant symptoms to HF diagnosis, particularly alignment with recommendations of the National Institute for Health and Care Excellence guideline for HF diagnosis.

Results 36 748 patients had a diagnosis of HF recorded that met the inclusion criteria between 1 January 2010 and 31 March 2013. For 29 113 (79.2%) patients, this was first recorded in hospital. In the 5 years prior to diagnosis, 15 057 patients (41.0%) had a primary care consultation with one of three key HF symptoms recorded, 17 724 (48.2%) attended for another reason and 3967 (10.8%) did not see their general practitioner. Only 24% of those with recorded HF symptoms followed a pathway aligned with guidelines (echocardiogram and/or serum natriuretic peptide test and specialist referral), while 44% had no echocardiogram, natriuretic peptide test or referral.

Conclusions Patients follow various pathways to the diagnosis of HF. However, few appear to follow a pathway supported by guidelines for investigation and referral. There are likely to be missed opportunities for earlier HF diagnosis in primary care.

https://doi.org/10.1136/heartjnl-2017-312183

Example

- NICE guideline for diagnosis of heart failure in primary care. GPs should order an echo and blood test and refer to a specialist within 6/52 of suspecting their pt has HF
- Q: how often does this happen?
- This could be framed in different ways…

Details

- GPs should order an echo and blood test and refer to a specialist within 6/52 of suspecting their pt has HF
- Should we try to identify that a GP did these things?
- What does specialist mean?
- What if only e.g. echo done?
- How do we know when they suspected HF? Which symptom codes? When recorded? They're common! What if no symptom codes recorded? Or is suspicion starting a loop diuretic?

Imperial College London

What we did and found

- Divided pts into whether they had coded symptoms
- Five-year lookback from HF dx used: “dx date” taken as first recorded HF code (80% in HES, 20% CPRD) – not perfect
- First symptom in those 5 years taken as pt’s presentation to primary care with (early) HF – not perfect either
- Full, partial and zero guideline compliance defined
- Only 7% full compliance in 6/52, 24% in 5 years

DATA CONSIDERATIONS

Imperial College London

General considerations

- Data Sources: electronic health records, registries, surveys, administrative datasets (local and national)
- Ethical Considerations: patient consent, IRAS (etc…), data protection regulations (e.g. small number suppression)
- Key Variables: demographics, diagnoses, treatments, metrics of policy success, policy change dates, confounders
- Data Cleaning: resolving duplicate entries (e.g. >1 date of death), choosing codes, managing missing values, ensuring longitudinal linkage (do pt IDs match?)
- Quality Checks: consistency (e.g. LTCs), accuracy (validation studies?), plausibility assessment (BP=20, BMI=100)
- Format for analysis: wide, long, aggregated

METHODOLOGICAL CHALLENGES

General methodological challenges

- Confounding: differences in patient characteristics, co-occurring interventions, secular trends
- Selection Bias: non-random policy adoption, varying implementation fidelity, attrition (but ITT)
- Measurement Error: inaccurate coding, misclassification of group or outcome
- Generalizability: applicability of findings to wider populations or settings (might not matter)
- Statistical Power: adequacy of sample size for detecting meaningful changes

2: CHOOSE THE METHOD OF ANALYSIS

Options we'll cover today

- Simple before vs after comparisons
- Difference in difference
- Interrupted time series
- Statistical process control inc funnel plots

Simple before v after comparison

- RQ: did GP compliance with HF dx guideline improve after publication of plan to improve CVD dx?
- Define 2 time points, e.g. first year after publication c.f. year before
- Choose metric(s): referral and blood test following pt presentation with possible HF symptom(s)
- Operationalise in the data with SNOMED codes etc
- Choose confounders – patient age, sex etc. Also ineq
- Analyse: chi-sq or logistic regression

Pros of this approach

- Simple – esp if no confounding is assumed
- Quick – used a lot in clinical audit
- That's it

Cons of this approach

- Highly susceptible to regression to the mean e.g. enrolment of pts w severe symptoms, HIUs
- No control for pre-existing trends
- No control for seasonality
- No control for external influences – can we attribute any change in the metrics to the “intervention” of interest?

Difference-in-difference (DID)

- Extends the before v after comparison to include a control group. Aka controlled before and after design
- The counterfactual is based only on the control group, thereby assuming parallel (pre-intervention) trends
- Testing this assumption needs more data than you might have

Example: centralisation of stroke services in England into hubs and spokes

- 2010 reorg: London (8 HASUs, 24 other), Manchester (3 HASUs, 11 other but v diff criteria and model), rest of England (no reorg)
- Q: did the reorg “work”?
- Controlled before-and-after design to compare sites pre and post centralisation (in terms of impact of centralisations on clinical outcomes, delivery of clinical interventions and cost-effectiveness) + wider comparisons with the rest of England

Some results for London

- Sig better than the RoE re mortality [–1.1%, 95% CI –2.1% to –0.1%; estimated 96 more lives saved per year], mean length of stay (–1.4 days, 95% CI –2.3 to –0.5 days)
- Analyses of data to March 2016: reductions in mortality and LOS were sustained, and delivery of clinical interventions either improved or sustained
- https://www.journalslibrary.nihr.ac.uk/hsdr/HSDR07070#scientific-summary for more

Interrupted time series (ITS)

- Use data at regular intervals, e.g. monthly guideline compliance rates, for intervention group only
- Explicitly model pre-intervention and post-intervention trends plus step change at t=0
- The pre-intervention trend serves as the counterfactual
- Need to check for autocorrelation, e.g. ACF plot and Durbin-Watson residuals: ARIMA, AR(1) etc if found
- Plot the modelled trends as easier to interpret than coefficients (but give them too)

Pros of this approach

- No control group -> no problems due to between-group differences e.g. selection bias, unmeasured confounding problems
- Modelling the underlying trend controls for within-group characteristics that tend to change only slowly over time, secular changes, random fluctuations from one time point to the next and regression to the mean

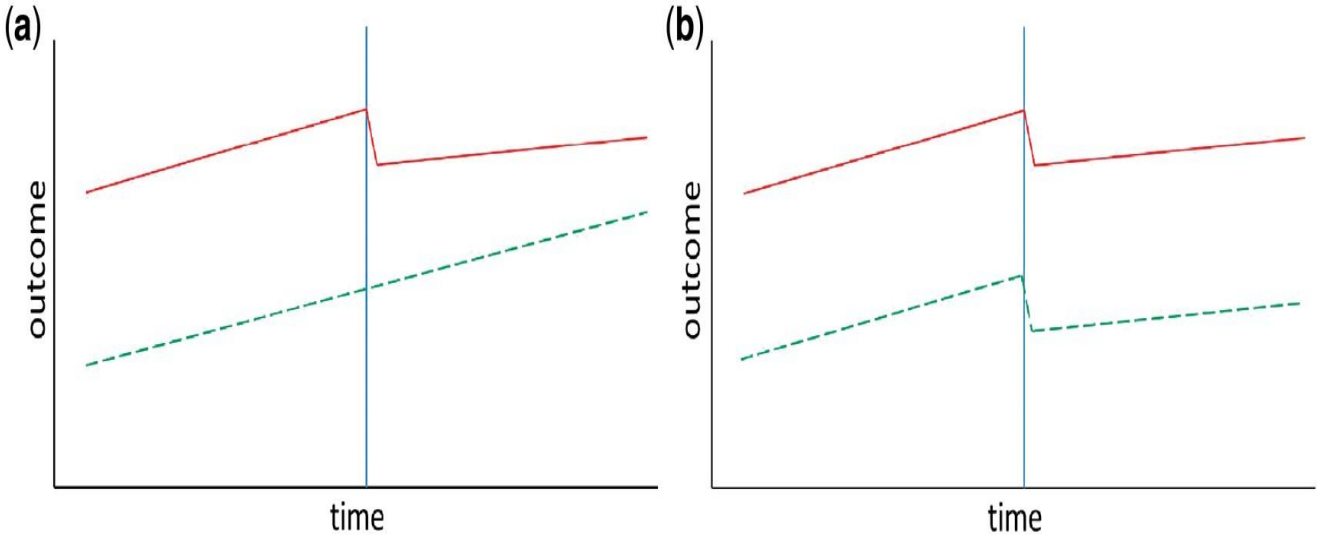
Cons of this approach

- More complex than before v after
- Can’t deal with other interventions / events occurring around the time of the intervention that may also affect the outcome of interest
- But… adding a control group (CITS) helps: if you see an effect in the intervention but not in the control group, it strengthens causal inference

How to analyse and interpret CITS

- Analyse two groups separately as ITS
- Then model together, inc an interaction, as CITS
- Compare the two – hopefully they agree!
- If only ITS shows intervention effect, then suspect simul event
- If only CITS shows intervention effect, then suspect change in control group due to some event affecting it only

Figure 1 Controlled interrupted time series. Solid line = intervention series, dashed line = control series: (a) effect in intervention group only; (b) suggests other event is at work



From: Int J Epidemiol, Volume 47, Issue 6, December 2018, Pages 2082–2093, https://doi.org/10.1093/ije/dyy135

Some thoughts on control groups

- A good idea if you can – RCTs do this
- Relevant unaffected groups may not exist, e.g. national roll-out / legislation, financial crisis: ITS is best here; also consider “control outcome” (one unaffected by the intervention)
- Ideal control is the same in terms of all variables other than exposure to the intervention
- Regression, PSM and synthetic controls all aim to do this with *measured* variables

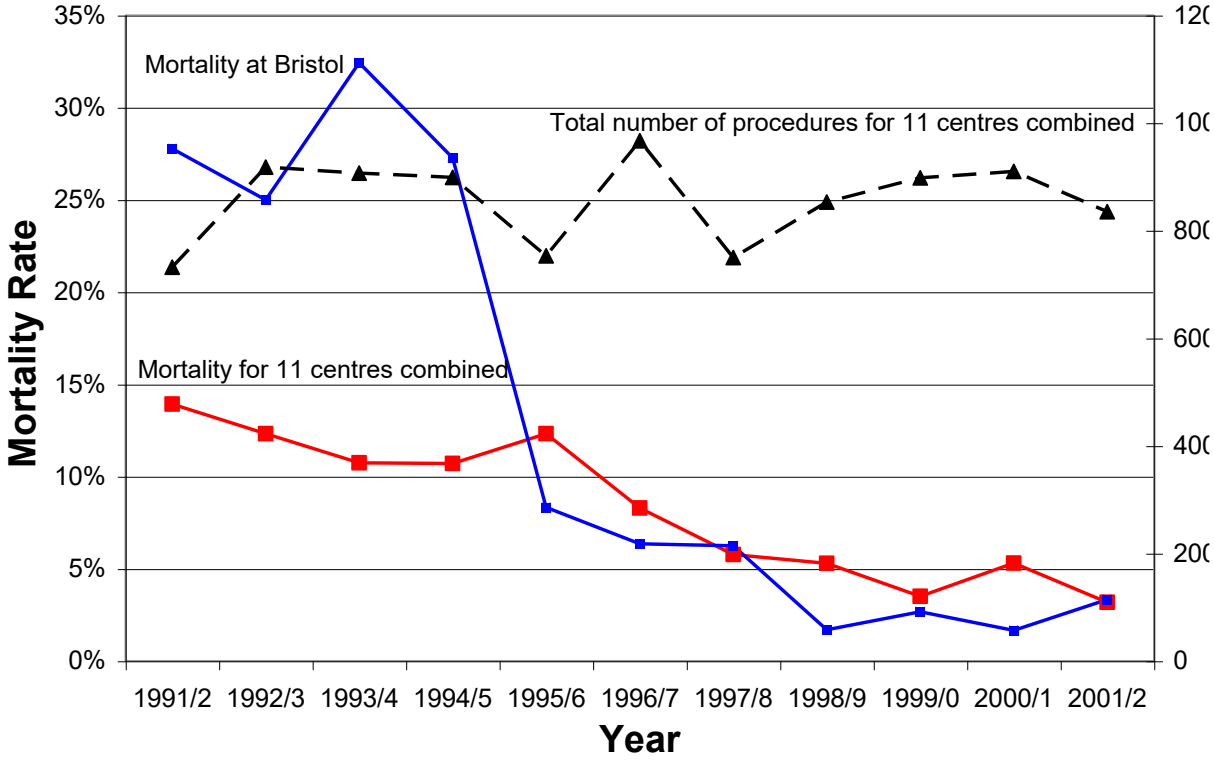
Some problems with control groups

- Don’t pick one that’s exposed to events to which the intervention group is not exposed – another reason why ITS is so popular
- Both groups need to have similar susceptibility to the intervention, so try to match (PSM or synthetic) pt factors throughout the study period, not just at t=0
- Contamination effects esp with behaviour change interventions e.g. control GPs hearing about it

Other things to consider with ITS

- It includes linear trend estimation, so need 5+ time points. Non-linear trends harder to do and interpret, so DID?
- If modelling counts, look for overdispersion. Model with quasi-Poisson, neg bin, gamma dist etc
- **Pre-specified** sensitivity analyses on control choice recommended
- COVID-19 impact: segment or exclude period

Check for unintended consequences: example of Bristol Inquiry – operations and mortality April 1991 to April 2002



STATISTICAL PROCESS CONTROL (SPC)

Introduction

- Created in 1920 at Bell Labs by Walter Shewhart to improve telephone equipment reliability; first control chart 1924
- He spread his methods to the US Army (arms manufacturing) and post-war Japan
- Sources of variation were “common” (an in-control process) or “special” (an out-of-control process)
- Test for deviations from expected performance
- It’s the standard approach for quality improvement

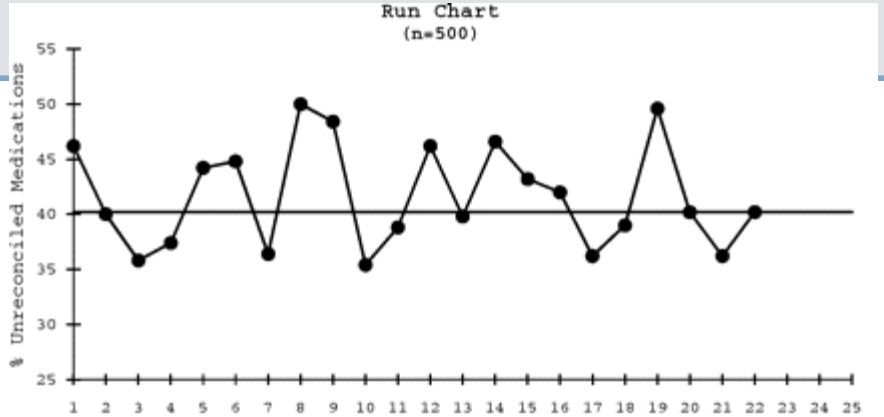
Imperial College London

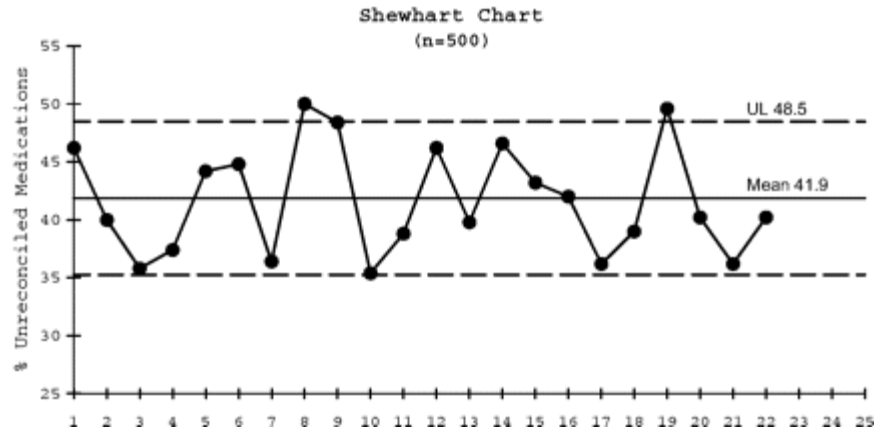
It avoids this error… (from "Making data count – an interactive guide")

The procedure

- Collect regular measurements (e.g., monthly guideline compliance rates, median waiting time)
- Calculate central line (mean or median) and control limits (typically ±3 sigma [SD of the process] except for run charts etc)
- Plot values and flag points outside control limits or unusual patterns
- Use rules to detect non-random variation (trend, shift, cycle)

Run chart and Shewhart chart compared (from Perla et al, BMJQS 2011)



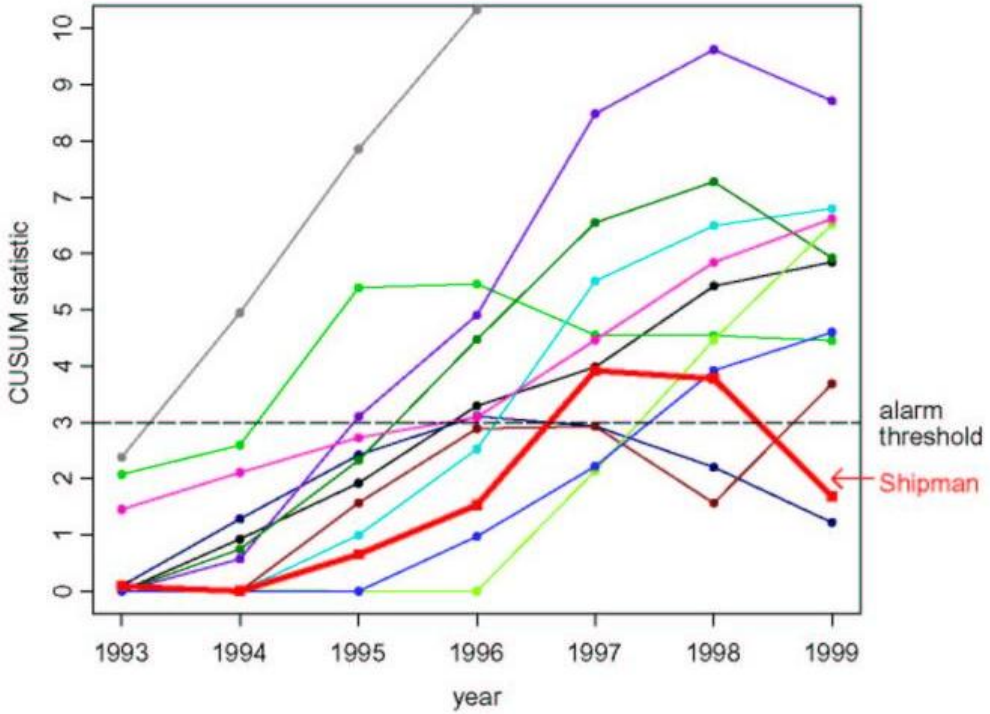


Imperial College London

Many types!

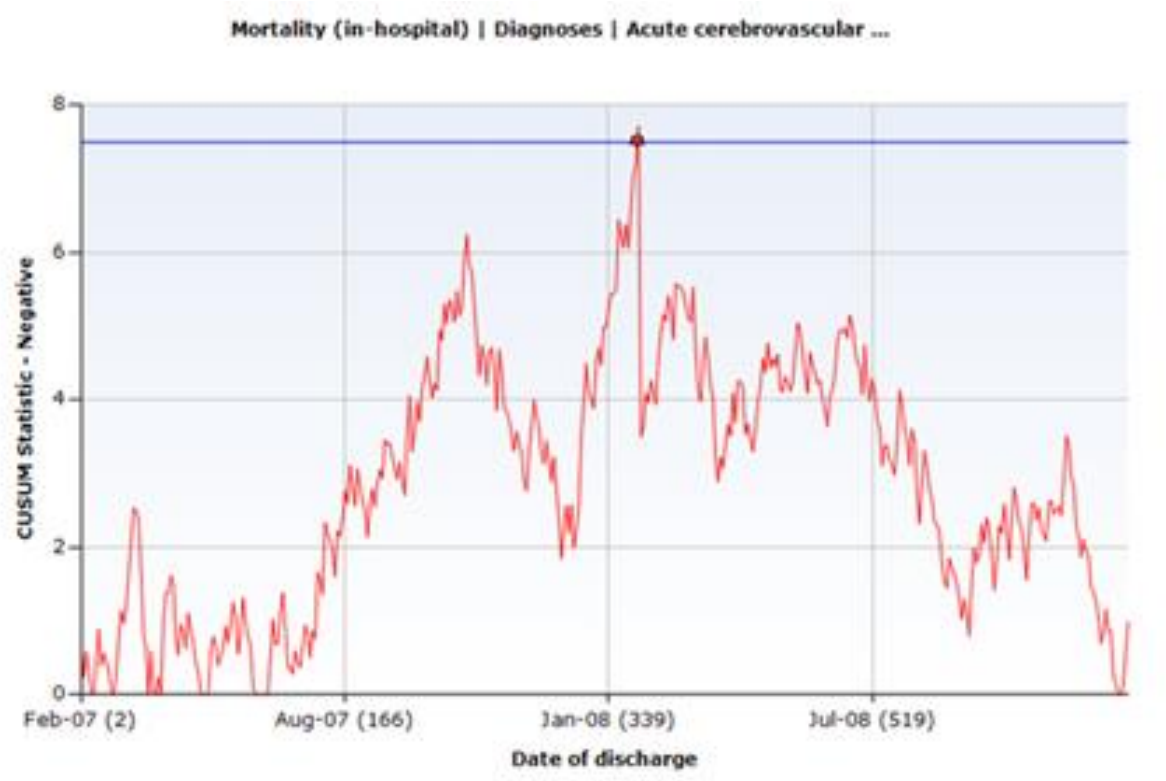
- Run charts – easy but can't spot special cause var
- Shewhart charts – good for big jumps only
- P charts – for proportions. If p is rare, try g charts
- XmR charts (individuals and moving range) – good if only one obs per time point (monthly infections, mean daily ambulance response times etc). Versatile
- CUSUM – good for small shifts
- Funnel plot – good for comparing multiple units against the same benchmark

CUSUM charts for the GPs that signalled as having unusually high mortality rates



Imperial College
London

Example of risk-adjusted CUSUM



Annals of Surgery • Volume 249, Number 3, March 2009

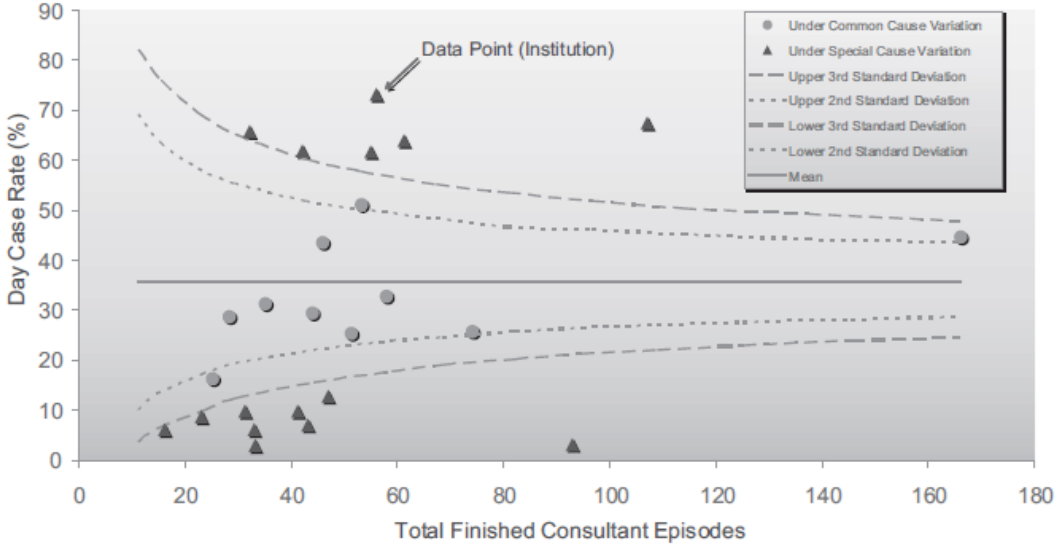


FIGURE 1. Funnel plot of day case rates for hemorrhoidectomy in London acute NHS trusts (2005/06). Source data from Hospital Episode Statistics online.[39]

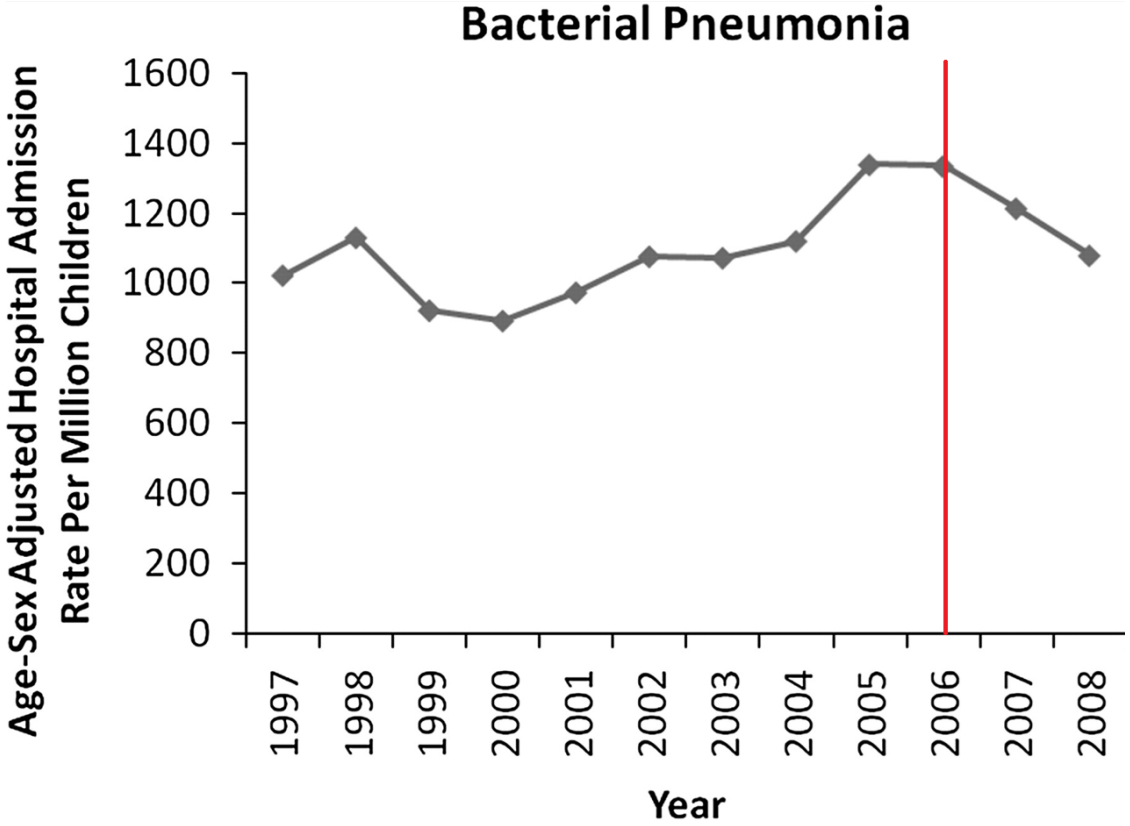
Some key considerations for control charts

- How many data points? Preferably 15+, else run chart
- When to recalculate centre line and limits? Only when sure a change has occurred, it's been sustained, and you've understood why the change happened
- What if you want to monitor multiple metrics? Choices are i) separate SPCs, ii) create composite metric, iii) multivariate SPC if correlated

Imperial College London

Quiz: what would you do with this?

PCV7 vaccination introduced Sep 2006. Did it reduce bacterial pneumonia admissions?



Join at menti.com | use code 5430 4217

Mentimeter

Menti

New presentation

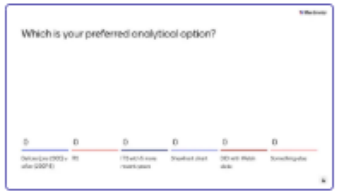
Which is your preferred analytical option?

menti.com

5430 4217

Waiting for participants

Select which slide to add



0	0	0	0	0	0
Before (pre-2005) v after (2007-8)	ITS	ITS with 5 more recent years	Shewhart chart	DID with Welsh data	Something else

Some thoughts on those options

- Before (pre-2006) v after (2007-8): we did this
- ITS: not enough post-intervention time points
- Shewhart chart: could base centre line on pre-intervention rate but not v sensitive
- DID with Welsh data: good if you have the data, assuming Wales didn't do same roll-out
- Or: use quarterly data but could be noisy. Or?

Concluding thoughts

- Approach is determined by how you frame the question, availability of data, suitable control group(s) and ease of interpretation
- The wrong control group or benchmark is worse than none
- SPC good esp for QI, but ITS best for effect size estimation and economic evaluation

Imperial College London

Contact details

- Thank you!
- Email robert.bottle@imperial.ac.uk
- Connect with me on LinkedIn