Inspired by research. Driven by compassion. Inspiré par la recherche. Guidé par la compassion.

STUDY DESIGNS FOR EVALUATING EFFECTIVENESS OF AUDIT AND FEEDBACK

International Audit and Feedback Summit
Toronto May 4, 2018

JEREMY GRIMSHAW
MONICA TALJAARD

Ottawa Hospital Research Institute and University of Ottawa

The Ottawa Hospital | L'Hôpital d'Ottawa

OUTLINE

1. Introduction – 5 min
2. Case study: BORN-MND – 5 min

   Audience participation 5 min
3. Considerations in choosing a study design – 10 min

   Audience participation 10 min
4. Randomized designs – 40 min

   Audience participation 15 min
5. Non-randomized designs – 10 min

   Audience participation 10 min
6. Case study: BORN-MND – 5 min
7. Future directions and wrap up – 5 min

OVERVIEW OF THE WORKSHOP

- This workshop is intended to be interactive
- We will introduce the BORN-MND study at the beginning of the workshop
- As we progress through the workshop, we will pause several times to allow you to discuss the material, in particular, to discuss how to design an evaluation of the BORN-MND intervention
- Crib-notes are provided!
- We will ask 1-2 tables to report back on their discussions
- At the end of the workshop, we will reveal the actual study design that was used

1. INTRODUCTION

CONTEXT

- Setting:
  - A&F being provided “in the real world”
- Interventions:
  - Embedded into existing QI programmes
  - Complex (multiple interacting components)
  - Delivered at the level of the provider or site (“cluster”)
- Outcomes:
  - Observed on multiple individuals (patients) per cluster
  - Usually obtained from routinely collected sources

PURPOSE OF EVALUATION

- Program evaluation
  - Addressing local question, did our program appear to achieve our aims
- Research evaluation
  - Addressing generalizable question, does audit and feedback work (it does, stop asking this question), how, when and why does audit and feedback work, how can we optimize audit and feedback within specific settings.
  - Research evaluation will (almost always) also address the local question
- Implications for design choices
  - May need less confidence about causality when undertaking program evaluation

KEY CONSIDERATIONS IN CHOOSING A STUDY DESIGN

- Can the delivery of the intervention be manipulated (i.e., can we use randomization)?
- How many independent providers/sites are available?
- Is there a requirement that the intervention be introduced at all sites (or can it be withheld from some sites)?
- Is it logistically feasible to introduce the intervention simultaneously across all sites?
- Are pre-intervention outcome data available to use in the evaluation?

2. CASE STUDY: BORN-MND

BMJ Quality & Safety Online First, published on 24 November 2017 as 10.1136/bmjqs-2017-007361

ORIGINAL RESEARCH

OPEN ACCESS

Effect of a population-level performance dashboard intervention on maternal-newborn outcomes

Deborah Weiss,[1] Sandra I Dunn,[1,2] Ann E Sprague,[1,2] Deshayne B Fell,[2,3] Jeremy M Grimshaw,[4] Elizabeth Darling,[5] Ian D Graham,[4] JoAnn Harrold,[6,7] Graeme N Smith,[8] Wendy E Peterson,[9] Jessica Reszel,[1,2] Andrea Lanes,[1,4,10] Mark C Walker,[1,3,4,10,11,12] Monica Taljaard[4]

Objectives To assess the effect of the Maternal Newborn Dashboard on six key clinical performance indicators in the province of Ontario, Canada.

Design [redacted] using population-based data from the provincial birth registry covering a 3-year period before implementation of the Dashboard and 2.5 years after implementation (November 2009 through March 2015).

Setting All hospitals in the province of Ontario providing maternal-newborn care (n=94).

Intervention A hospital-basedonlinee audit and feedback programme.

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2017-007361).

For numbered affiliations see end of article.

Correspondence to

long-term health of women and infants.[1] Across Canada, there is wide variability in clinical practice and outcomes in maternal-newborn care settings, which suggests there are opportunities for improvement.[2 3] One approach that has been widely used to promote evidence-based care in clinical settings is audit and feedback,[4] in which clinical performance is assessed over time and feedback

SETTING

- BORN Ontario launched January 2012: “The Better Outcomes Registry & Network” (a provincial birth registry)
- Involves all hospitals in Ontario providing maternal newborn care
- November 2012: BORN launched an A&F intervention, called the Maternal Newborn Dashboard (MND)
- MND embedded into the data collection process for the Registry
- Population-level data available for all N=96 maternal newborn hospitals in Ontario from 2009-2015

INTERVENTION

- Maternal Newborn Dashboard (MND)
  - Near real-time feedback on 6 Key Performance Indicators (KPIs)
  - Compares performance to established benchmarks
  - Compares performance to peers
  - Provides alerts when performance is sub-optimal
  - Provides evidence summaries

INTERVENTION

- Maternal Newborn Dashboard (MND)

Key Performance Indicators	Rate (%)	Status	Benchmark values (%)			Comparator values (%)		
			Target (green)	Warning (yellow)	Alert (red)	Other Neonatal Level IIc hospitals	Other 1001-2499 birth volume hospitals	Ontario
1 Proportion of newborn screening samples that were unsatisfactory for testing	1.2		<2.0	2.0-3.0	>3.0	1.1	1.5	1.1
2 Rate of episiotomy in women who had a spontaneous vaginal birth	12.3		<13.0	13.0-17.0	>17.0	15.6	10.0	11.2
3 Rate of formula supplementation at discharge in term infants whose mothers intended to breastfeed	35.6		<20.0	20.0-25.0	>25.0	34.0	33.6	32.7
4 Proportion of women with a cesarean section performed from ≥37 to <39 weeks' gestation among low-risk women having a repeat cesarean section at term	42.3		<11.0	11.0-15.0	>15.0	45.8	48.0	41.1
5 Proportion of women who delivered at term and had Group B Streptococcus (GBS) screening at 35-37 weeks' gestation	90.2		>94.0	90.0-94.0	<90.0	92.3	88.7	91.4
6 Proportion of women who were induced with an indication of post-dates and were less than 41 weeks' gestation at delivery	17.2		<5.0	5.0-10.0	>10.0	22.6	27.4	19.1

Target Alert Warning

AUDIENCE PARTICIPATION – 5 MIN

- How should we evaluate the effectiveness of the MND implementation?
  - Consider the “Key considerations in choosing a study design” with respect to the MND evaluation
  - Discuss possible ways to evaluate the BORN MND intervention

3. CHOOSING A STUDY DESIGN

- General principle:
  - Choose the most robust design possible to minimize bias while maximizing generalizability
- Minimizing bias (internal validity)
  - Is the observed improvement actually caused by the A&F?
- Maximizing generalizability (external validity)
  - Will the A&F also work in other sites/providers and other patients?

TWO MAIN TYPES OF STUDY DESIGNS

- Randomized controlled trials
- Non-randomized (Quasi-experimental) studies

EVALUATING A&F

- To evaluate effectiveness of an intervention, we need a comparator
- Examples:
  - A&F versus no A&F (not ideal)
  - Usual A&F versus new A&F
  - A&F + something else versus A&F alone

RANDOMIZED DESIGNS

- "Randomized controlled trial" (RCT)
  - Allocate an adequate number of independent units (e.g., sites, providers) to different interventions ("study arms") using a random procedure (preferably computer-generated)
  - Randomization serves to "equalize" the groups being compared
  - Differences observed between the study arms can be confidently attributed to the intervention
- Randomized designs always preferable

NON-RANDOMIZED DESIGNS

- "Observational" or "Quasi-experimental" design
  - Non-random distribution of sites / providers across the study arms (e.g., based on own preferences, logistical considerations)
- Differences observed cannot be attributed to the intervention without making some strong assumptions
- Should be used only when no other choice, e.g.:
  - All providers/sites must receive intervention at the same time
  - Only a small number of providers/sites available (not enough to randomize)

UNIT OF RANDOMIZATION

- Two types of randomized controlled trials:
  - Patient randomized trial
  - Cluster randomized trial (CRT)

- Patient randomization generally preferable (but not possible for site- or provider-level interventions such as A&F)

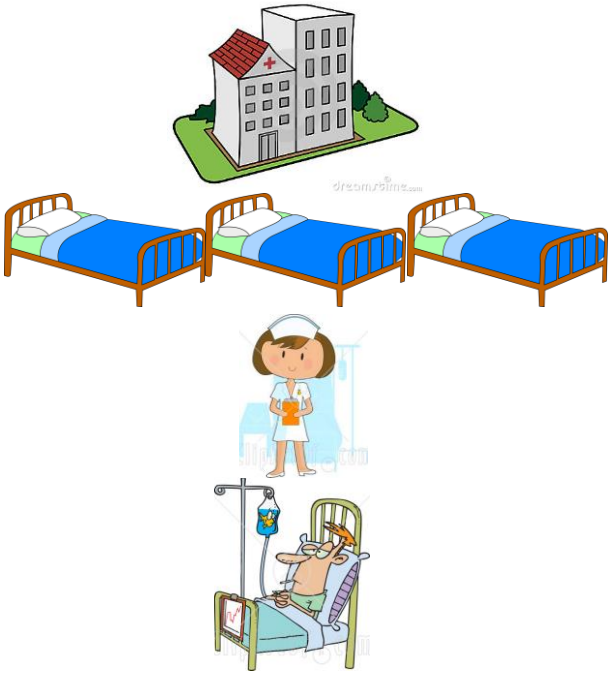
WHAT IS A CRT?

- A randomized trial in which intact groups (“clusters” of individuals, rather than separate individuals) are allocated to different study arms while outcomes are then observed on individuals within each cluster
  - Examples of clusters: providers, hospitals, nursing homes, primary care practices

- Key characteristic of a CRT:
  - Unit of randomization $\neq$ Unit of observation

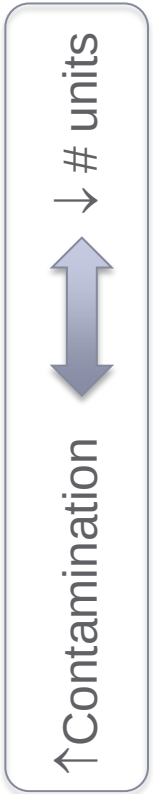
UNIT OF RANDOMIZATION

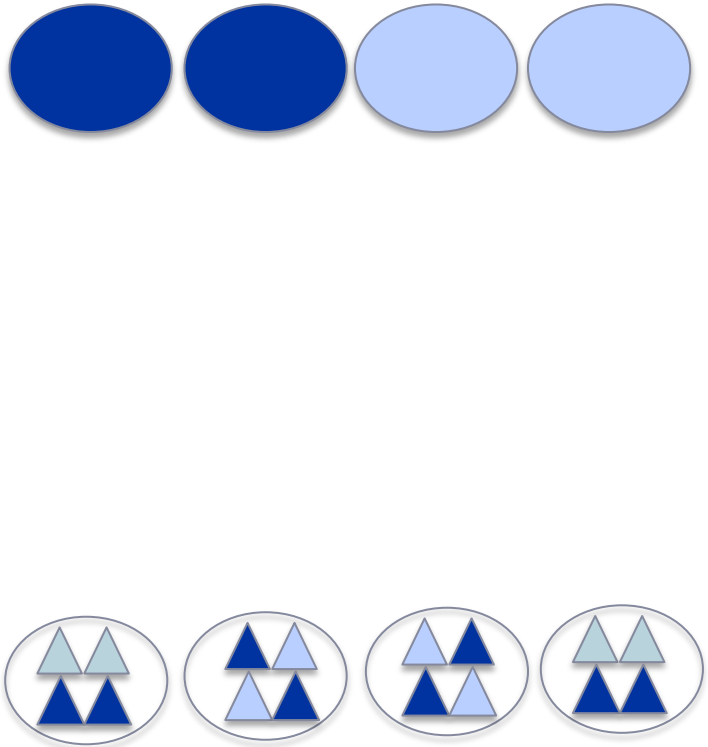
- At what level should we randomize:
  - LHINs (Local health networks)?
  - Individual hospitals?
  - Wards within hospitals?
  - Individual providers?
  - Patients?



RANDOMIZATION UNIT TRADE-OFFS

- Local health networks?
- Individual hospitals?
- Wards within hospitals?
- Individual health professionals?
- Patients?





KEY IMPLICATION OF CLUSTER RANDOMIZATION

- Responses of individuals in the same cluster usually more similar than responses of individuals in different clusters
  - Degree of similarity is measured by the “Intracluster Correlation Coefficient”
- Standard statistical methods assume observations are uncorrelated
- Adjustments to standard methods for sample size calculation and analysis are required
- Need to work with a statistician who is experienced in cluster randomized trials

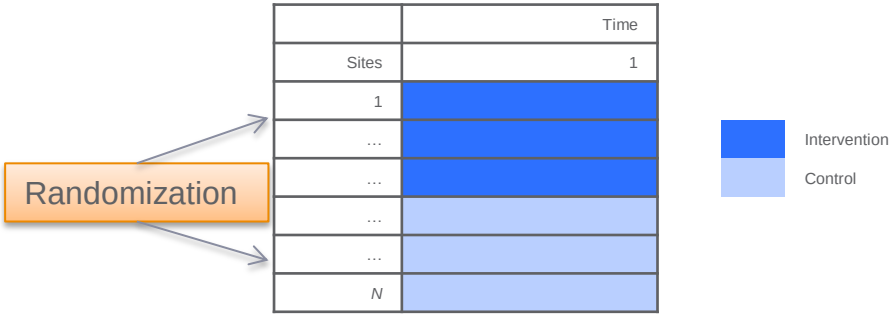
AUDIENCE PARTICIPATION – 10 MIN

- How should we evaluate the effectiveness of the MND implementation?
  - Consider the two main types of study design with respect to the MND evaluation
  - Discuss possible ways to evaluate the BORN MND intervention

4. RANDOMIZED DESIGNS

- Main cluster randomized trial (CRT) designs:
  1. Two arm parallel design
  2. Multi-arm parallel design
  3. Parallel arm before and after design
  4. Repeated measures parallel arm design
  5. Stepped wedge design
  6. Factorial trial design

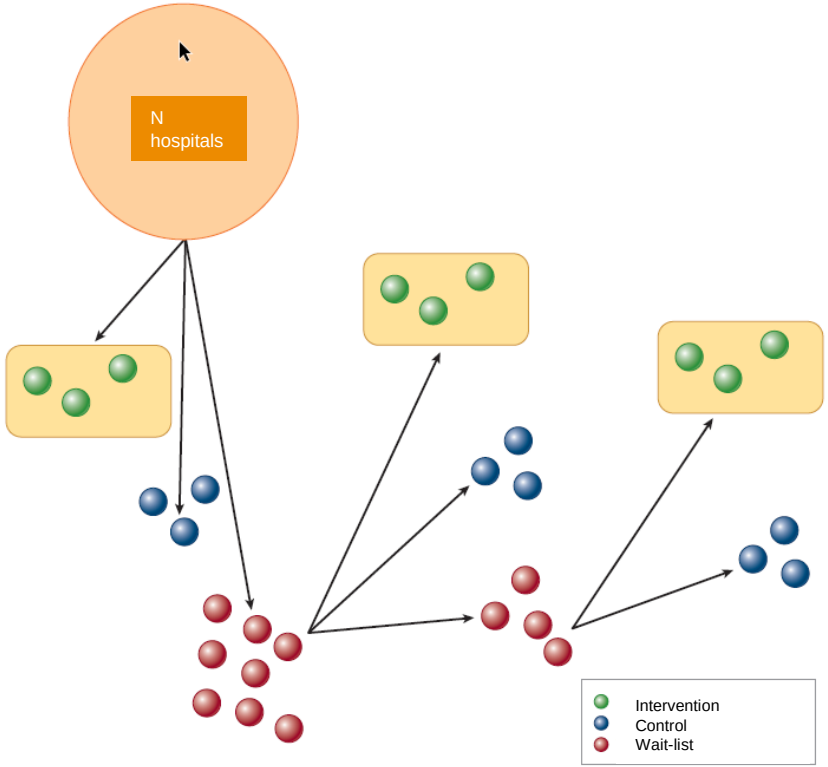
1. PARALLEL ARM DESIGN



- Advantages:
  - Simple to understand
  - Straightforward analysis
- Disadvantages:
  - Other more powerful designs are available (“power” = ability to detect an intervention effect)
  - Cannot assess baseline comparability in performance

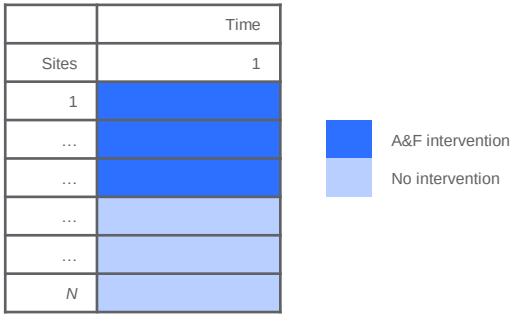
1. PARALLEL ARM WITH STAGGERED IMPLEMENTATION

- Simultaneous implementation of the intervention at many sites may be logistically challenging
- An alternative is to randomly allocate sites “in waves”



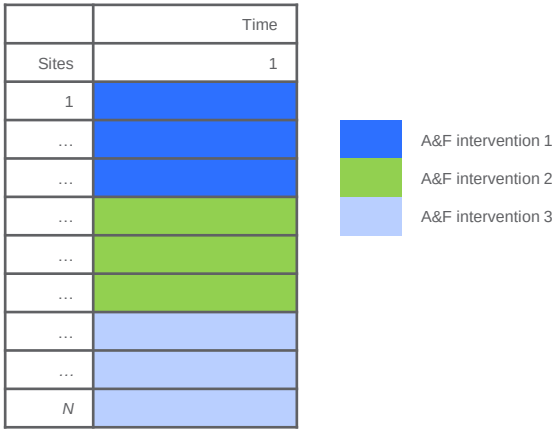
2. PARALLEL MULTI-ARM DESIGNS

- Two arms



Does it work?

Multiple arms

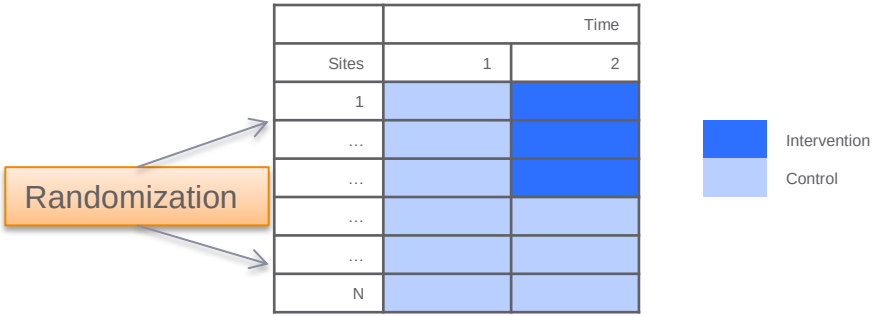


Which version works the best?

2. PARALLEL MULTI-ARM DESIGNS

- Advantages
  - Allows comparison of multiple interventions or levels of intervention under similar circumstances
- Disadvantages
  - Need more sites to achieve the same power (due to use of multiple arms)
  - Small differences between arms implies larger sample sizes required
  - Analysis more complicated (need to account for multiple comparisons)

3. BEFORE AND AFTER PARALLEL ARM



- Add a pre-intervention measurement in both arms

3. BEFORE AND AFTER PARALLEL ARM

- Advantages
  - Can assess whether sites in different arms are comparable before intervention
  - Utilizing the pre-intervention data in analysis can increase power
  - Can assess whether sites who are dropped from the analysis (e.g., due to closures, mergers, attrition) are similar to those who remain
- Disadvantages:
  - More complex analysis
  - Different methods of analysis are possible which may give different answers
  - May extend the total study duration if no routine data available

4. LONGITUDINAL PARALLEL ARM

- Multiple measurements taken
  - Before intervention and/or
  - During intervention and/or
  - After intervention

4. LONGITUDINAL PARALLEL ARM

A. Simple parallel arm

	Time
Site	1
1	
…	
…	
…	
…	
K	

Intervention

Control

B. Parallel arm repeated measures

	Time					
Site	1	2	3	4	5	6
1						
…						
…						
…						
…						
K						

C. Parallel arm before and after

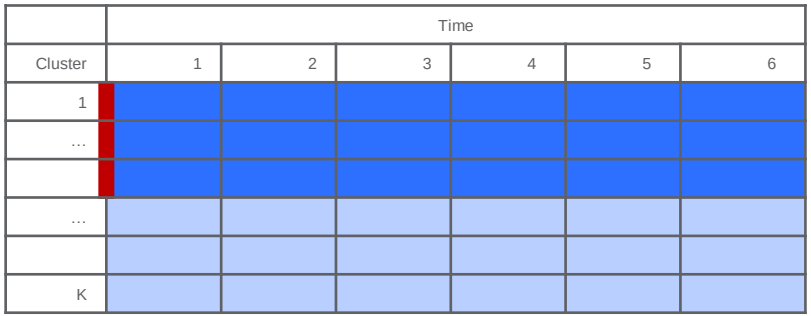
	Time	
Site	1	2
1		
…		
…		
…		
…		
K		

D. Parallel arm before & after repeated measures

	Time											
Site	1	2	3	4	5	6	7	8	9	10	11	12
1												
…												
…												
…												
…												
K												

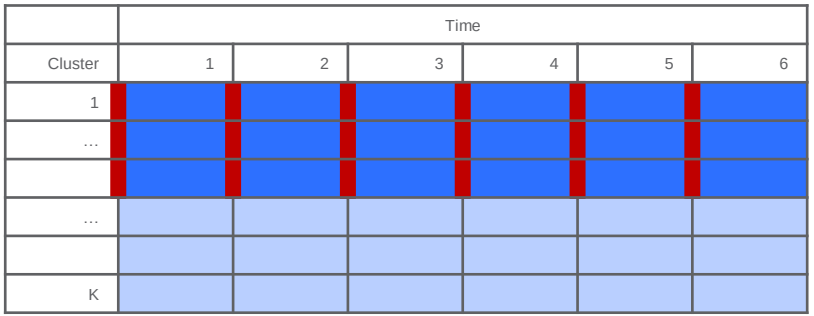
ISSUES TO CONSIDER IN LONGITUDINAL DESIGNS

Single delivery



Intervention has an immediate effect which is sustained over time

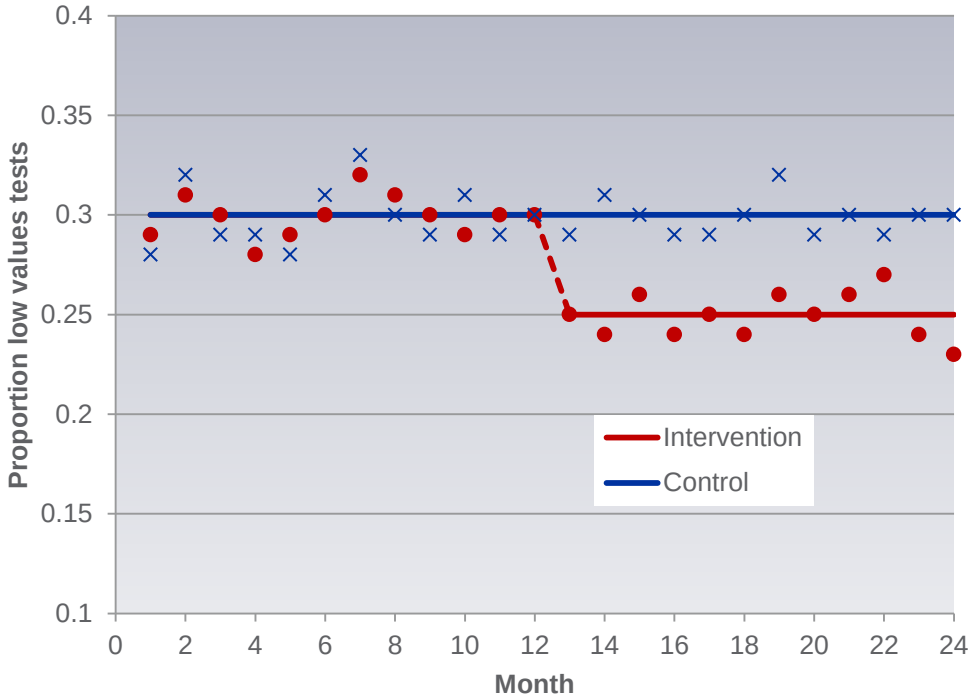
Repeated delivery



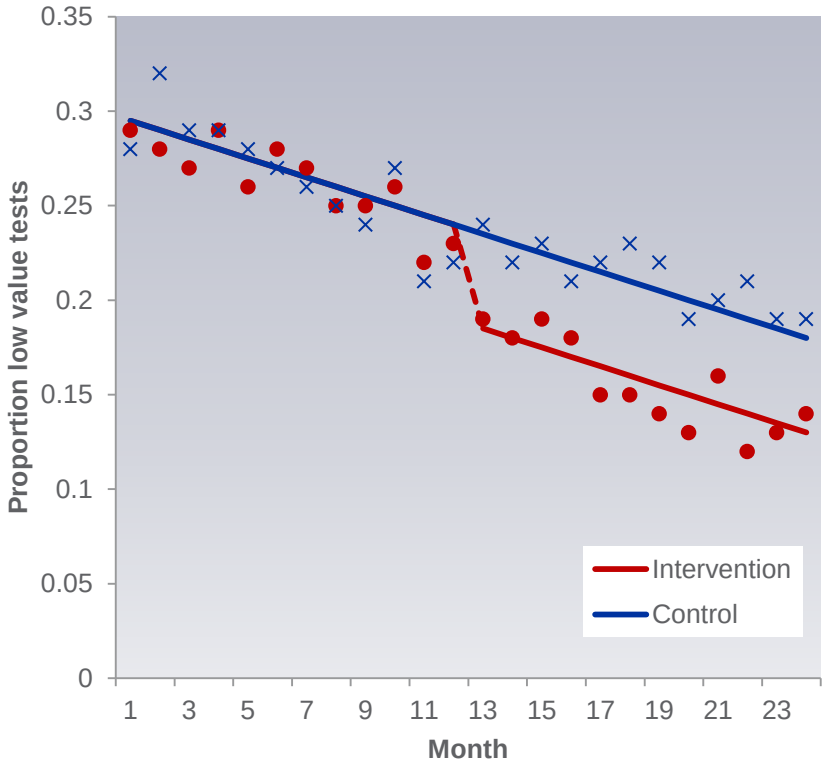
May need repeated delivery to ensure effect is sustained

ISSUES TO CONSIDER IN LONGITUDINAL DESIGNS

- Immediate change that persists through time



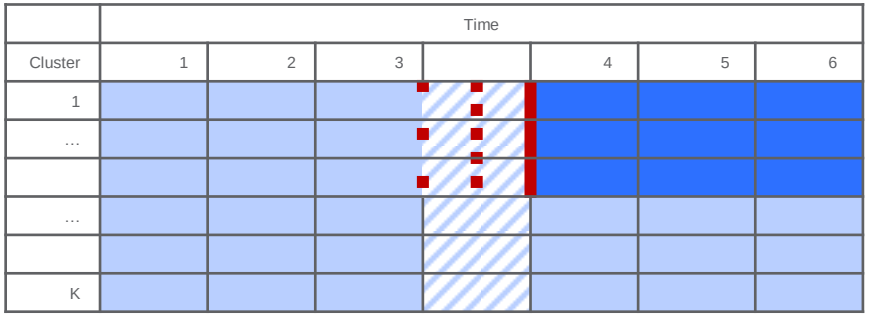
ISSUES TO CONSIDER IN LONGITUDINAL DESIGNS



- Immediate change on top of a secular trend
  - Outcomes already improving even before intervention
  - Intervention has an additional effect over and above the secular trend

ISSUES TO CONSIDER IN LONGITUDINAL DESIGNS

Allowing for an implementation period

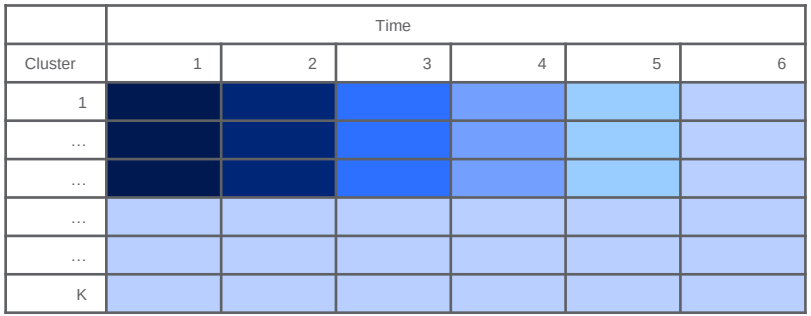


May need to allow for implementation period, or a delay before any effect can be observed.

During the implementation period, the site cannot be considered fully exposed to the intervention.

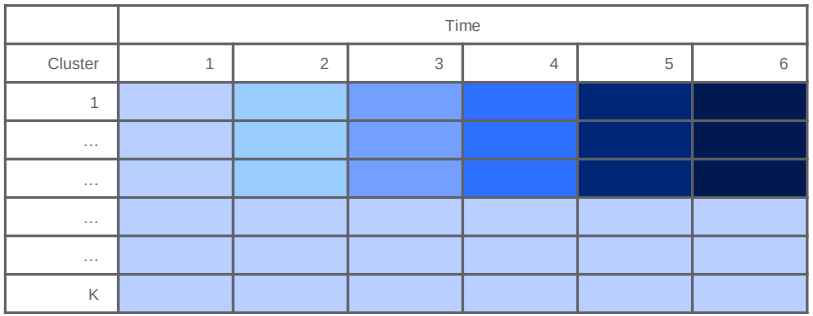
ISSUES TO CONSIDER IN LONGITUDINAL DESIGNS

Decay effects



Intervention has an immediate effect which decays over time

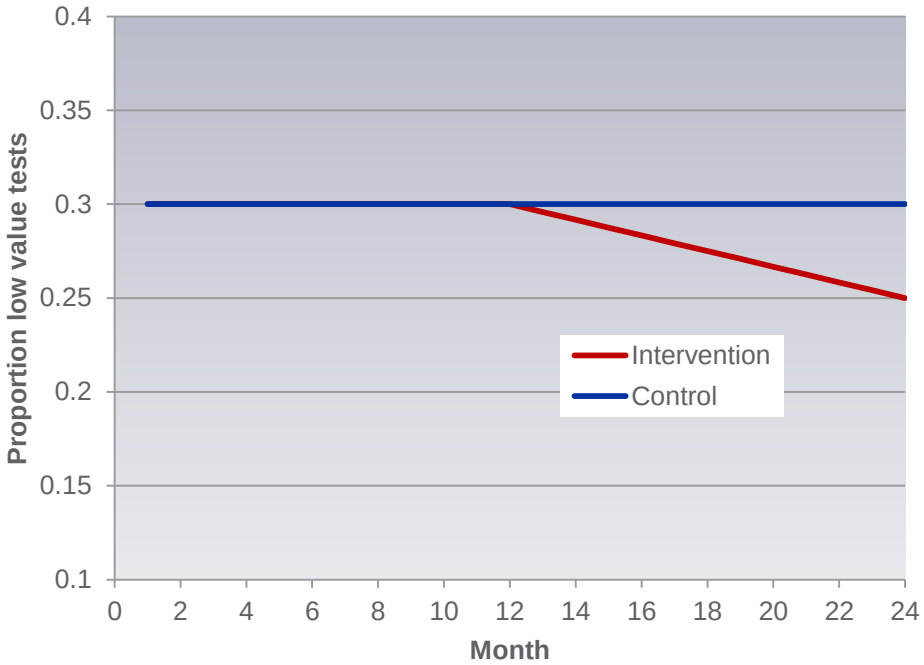
Learning effects



Intervention has a gradual effect

ISSUES TO CONSIDER IN LONGITUDINAL DESIGNS

- Gradual change



4. LONGITUDINAL PARALLEL ARM

- Advantages:
  - Can study how outcomes change over time in response to intervention (learning, decay)
  - Can assess whether changes are sustained in the long-run
  - Can assess for presence of “secular trends” (improvements happening naturally over time)
  - Can increase power
  - Can check baseline comparability in level and secular trend
- Disadvantages:
  - Can take longer to complete the study
  - May increase the risk of attrition
  - May increase the risk of contamination between arms
  - More complicated to analyze
  - Different methods of analysis can give different answers
  - Need a good understanding of how the intervention works

4. LONGITUDINAL TRIAL EXAMPLE

Data feedback and behavioural change intervention to improve primary care prescribing safety (EFIPPS): multicentre, three arm, cluster randomised controlled trial

Bruce Guthrie,[1] Kimberley Kavanagh,[2] Chris Robertson,[2] Karen Barnett,[3] Shaun Treweek,[4] Dennis Petrie,[5] Lewis Ritchie,[6] Marion Bennie[7,8]

[1]Population Health Sciences Division, University of Dundee, Dundee DD2 4BF, UK
[2]Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK
[3]Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK
[4]Health Services Research Unit, University of Aberdeen, Aberdeen, UK
[5]Centre for Health Policy, University of Melbourne, Melbourne, Australia
[6]Department of Academic Primary Care, University of Aberdeen, Aberdeen, UK
[7]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK
[8]Information Services Division,

ABSTRACT

OBJECTIVE
To evaluate the effectiveness of feedback on safety of prescribing compared with moderately enhanced usual care.

DESIGN
Three arm, highly pragmatic cluster randomised trial.

SETTING AND PARTICIPANTS
262/278 (94%) primary care practices in three Scottish health boards.

INTERVENTIONS
Practices were randomised to: "usual care," consisting of emailed educational material with support for searching to identify patients (88 practices at baseline, 86 analysed); usual care plus feedback on practice's high risk prescribing sent quarterly on five occasions (87 practices, 86 analysed); or usual care plus the same feedback incorporating a behavioural change component (87 practices, 86 analysed)

RESULTS
In the primary analysis, high risk prescribing as measured by the primary outcome fell from 6.0% (3332/55 896) to 5.1% (2845/55 872) in the usual care arm, compared with 5.9% (3341/56 194) to 4.6% (2587/56 478) in the feedback only arm (odds ratio 0.88 (95% confidence interval 0.80 to 0.96) compared with usual care; P=0.007) and 6.2% (3634/58 569) to 4.6% (2686/58 582) in the feedback plus behavioural change component arm (0.86 (0.78 to 0.95); P=0.002). In the pre-specified secondary analysis of change in trend within each arm, the usual care educational intervention had no effect on the existing declining trend in high risk prescribing. Both types of feedback were associated with significantly more rapid decline in high risk prescribing after the intervention compared with before.

CONCLUSIONS
Feedback of prescribing safety data was effective at reducing high risk prescribing. The intervention would

4. LONGITUDINAL TRIAL EXAMPLE

- **Objective:** Evaluate effectiveness of feedback on safety of prescribing in primary care
- **Design:** Three arm CRT involving 262 primary care practices in Scotland with repeated quarterly pre and post measurements over 3 years
- **Interventions:** (1) Usual care; (2) Feedback on high risk prescribing sent quarterly on five occasions; (3) Feedback plus behavioural change component
- **Primary outcome**: Composite of six prescribing measures relating to high risk use of antipsychotics, non-steroidal anti-inflammatories, and antiplatelets
- **Primary analysis:** Between-arm comparison in the final quarter (at the end of the trial). Secondary: Between-arm comparison of slope changes

4. LONGITUDINAL TRIAL EXAMPLE

- **Results**: High risk prescribing declined in all three arms, but intervention arms had significantly more rapid decline after intervention

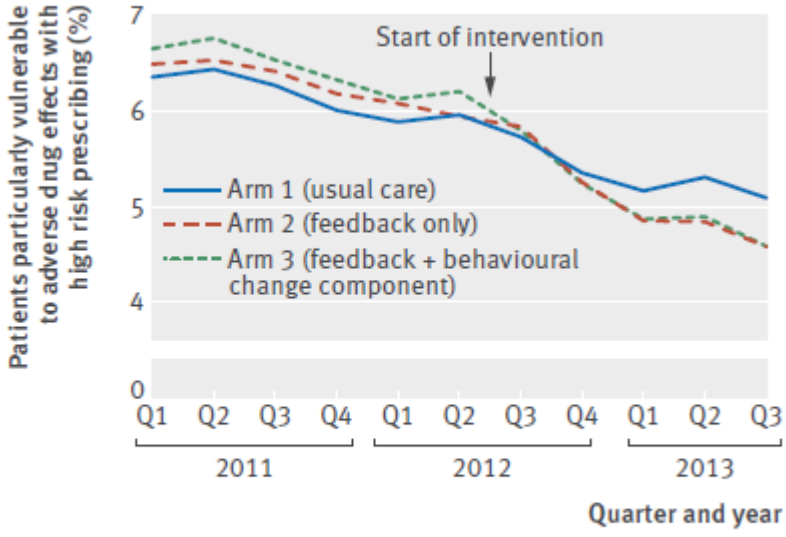
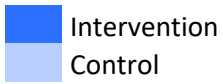


Fig 2 | Change in composite high risk prescribing in three trial arms

5. STEPPED WEDGE

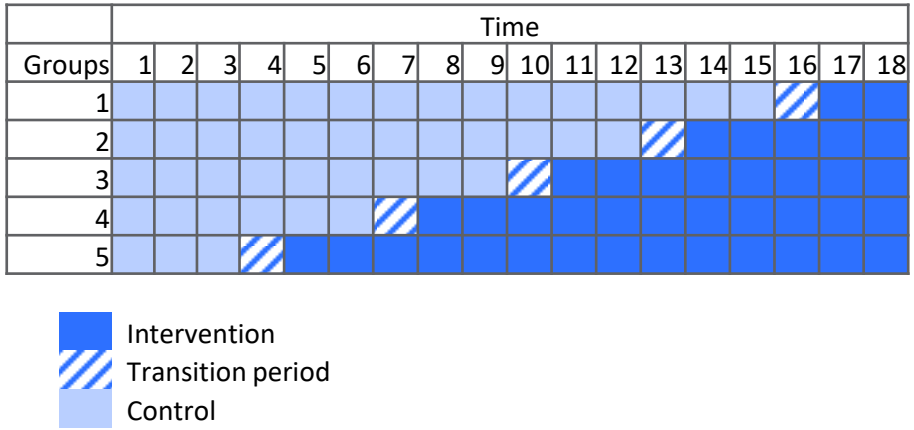
	Time					
Groups	1	2	3	4	5	6
1						
2						
3						
4						
5						



- All sites start in control and end in intervention condition
- Sites cross to intervention sequentially and in random order
- Outcomes are assessed repeatedly in each site over time

5. STEPPED WEDGE WITH TRANSITION PERIOD

- Can allow for a short transition period to allow the intervention to be put in place



5. STEPPED WEDGE: ADVANTAGES

- All sites receive the intervention during the study
- Uses randomization – better than implementing the intervention at all sites without any randomization
- May increase power over parallel arm designs
- Delivery of the intervention can be spread out over time (e.g., by having only one site or a small number of sites cross over each time)

5. STEPPED WEDGE: DISADVANTAGES

- All sites must be ready to implement intervention at any time
- Can increase the total duration of the study (increase risk that external events may influence outcomes)
- Some sites have to wait a long time before receiving intervention
- Heavy data collection burden (unless using routinely collected data)
- More complex to analyze and interpret results (can be difficult to separate the effect of the intervention from the effect of secular trends)

5. STEPPED WEDGE EXAMPLE

OPEN ACCESS Freely available online

PLOS ONE

The Feedback Intervention Trial (FIT) — Improving Hand-Hygiene Compliance in UK Healthcare Workers: A Stepped Wedge Cluster Randomised Controlled Trial

Christopher Fuller[1], Susan Michie[2], Joanne Savage[1], John McAteer[2], Sarah Besser[1¤a], Andre Charlett[3], Andrew Hayward[1], Barry D. Cookson[3], Ben S. Cooper[3¤b], Georgia Duckworth[3], Annette Jeanes[4], Jenny Roberts[5], Louise Teare[6], Sheldon Stone[1]*

1 Royal Free Campus, University College London Medical School, University College, London, United Kingdom, **2** University College London, London, United Kingdom, **3** Health Protection Agency, London, United Kingdom, **4** University College London Hospitals, London, United Kingdom, **5** London School of Hygiene and Tropical Medicine, London, United Kingdom, **6** Mid-Essex NHS Trust, Chelmsford, United Kingdom

Abstract

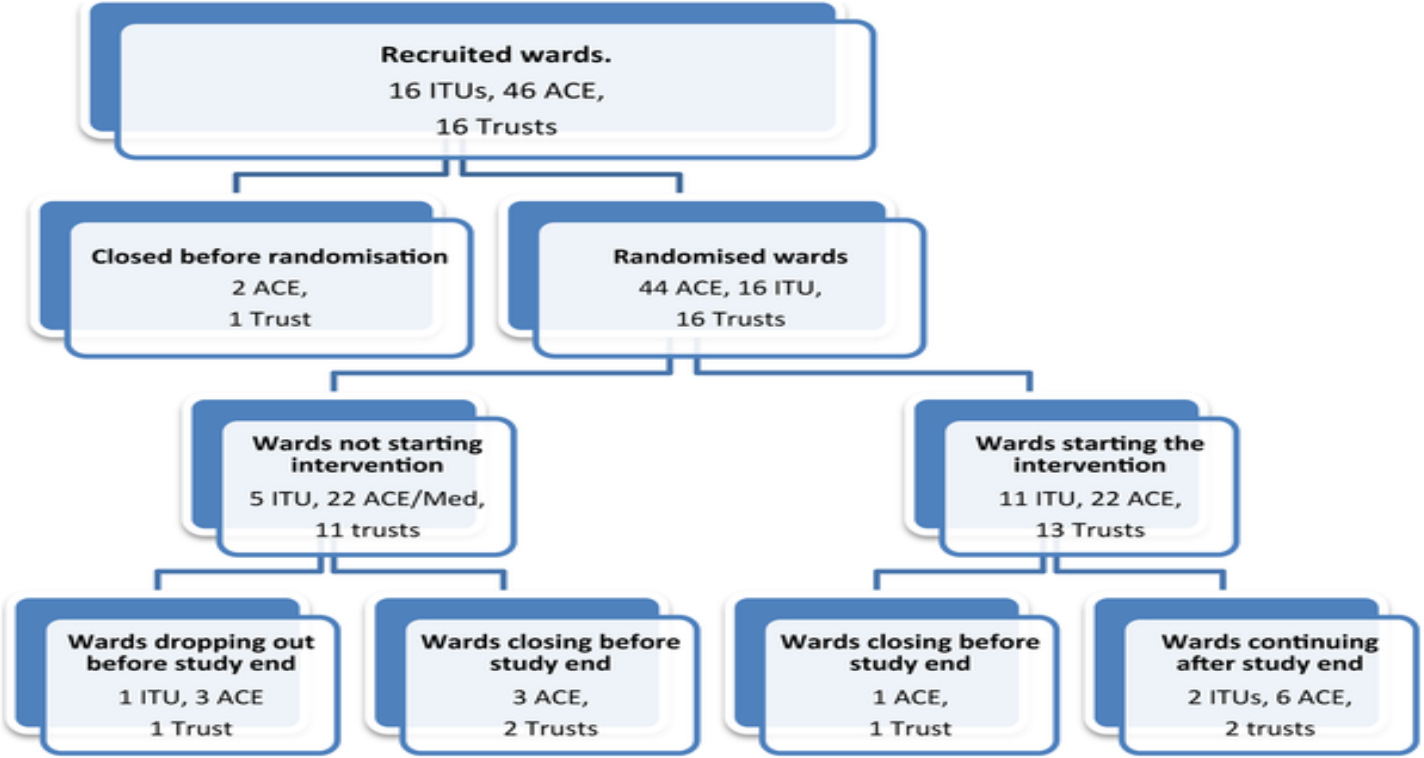
Introduction: Achieving a sustained improvement in hand-hygiene compliance is the WHO's first global patient safety challenge. There is no RCT evidence showing how to do this. Systematic reviews suggest feedback is most effective and call for long term well designed RCTs, applying behavioural theory to intervention design to optimise effectiveness.

Methods: Three year stepped wedge cluster RCT of a feedback intervention testing hypothesis that the intervention was more effective than routine practice in 16 English/Welsh Hospitals (16 Intensive Therapy Units [ITU]; 44 Acute Care of the Elderly [ACE] wards) routinely implementing a national cleanyourhands campaign). Intervention-based on Goal & Control theories. Repeating 4 week cycle (20 mins/week) of observation, feedback and personalised action planning, recorded on forms. Computer-generated stepwise entry of all hospitals to intervention. Hospitals aware only of own allocation. Primary outcome: direct blinded hand hygiene compliance (%).

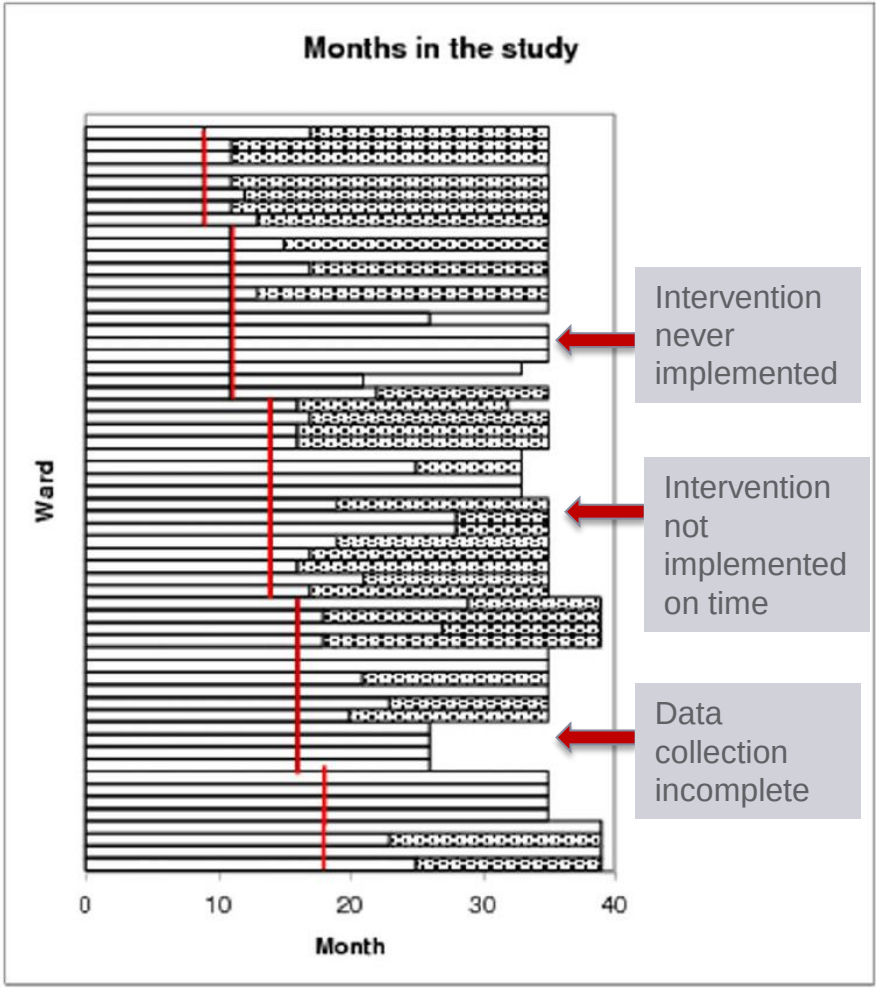
5. STEPPED WEDGE EXAMPLE

- **Objective:** To evaluate a behaviourally designed and theory-informed A&F intervention to improve hand-hygiene compliance
- **Design:** Three year stepped wedge CRT at 16 English/Welsh Hospitals; hospitals were randomized at 2 monthly intervals
- **Control:** Routine implementation of a national “cleanyourhands” campaign (consisting of bedside placement of alcohol hand-rub, posters, plus audit and feedback of hand-hygiene compliance at least once every 6 months)
- **Intervention:** Repeated 4 week cycles of observation, feedback and personalised action planning
- **Primary outcome:** Directly observed blinded hand hygiene compliance
- **Results:** Moderate but significant improvements in hand-hygiene compliance

Figure 1. Flowchart showing study recruitment and attrition.

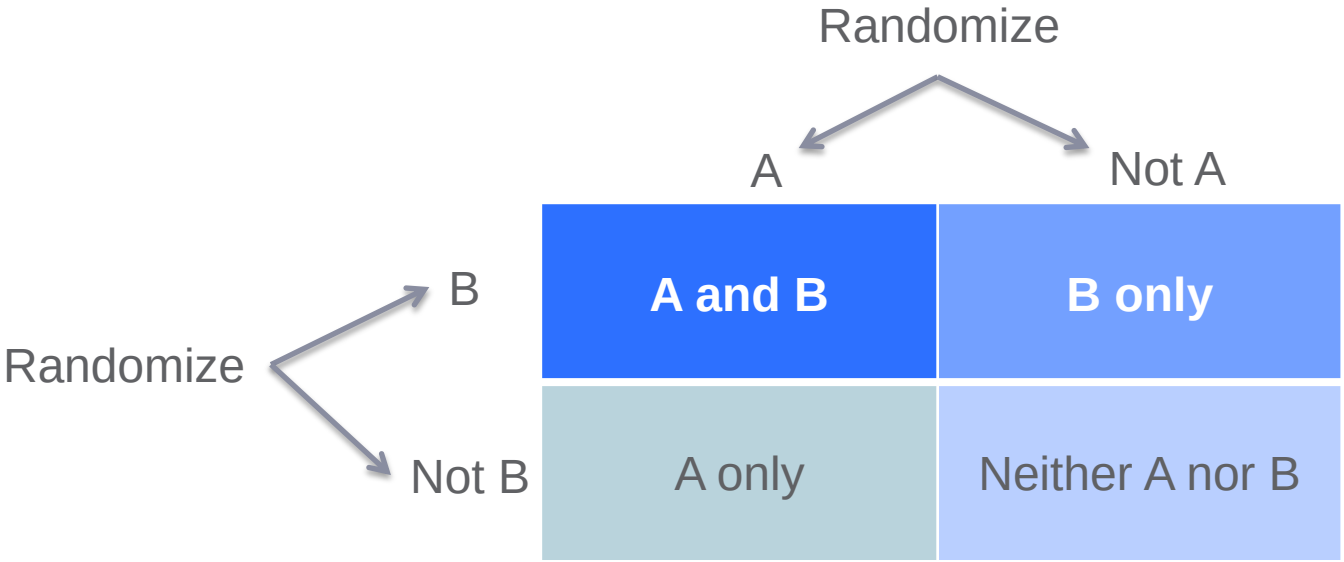


Fuller C, Michie S, Savage J, McAteer J, Besser S, et al. (2012) The Feedback Intervention Trial (FIT) — Improving Hand-Hygiene Compliance in UK Healthcare Workers: A Stepped Wedge Cluster Randomised Controlled Trial. PLOS ONE 7(10): e41617. https://doi.org/10.1371/journal.pone.0041617
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0041617


Months in the study
Ward
0
10
20
30
40
Month
Intervention never implemented
Intervention not implemented on time
Data collection incomplete

6. FACTORIAL DESIGN

- 2x2 factorial design



6. FACTORIAL DESIGN: INTERACTIONS

- Factorial design works best when there is no interaction
- No interaction:
  - Effect of each intervention is the same, regardless of whether the other is present or absent
- Interaction:
  - Effect of each intervention is different in the presence or absence of the other
    - Antagonistic: Effect of both interventions combined is smaller than the sum of their separate effects
    - Synergistic: Effect of both interventions combined is larger than the sum of their separate effects

6. FACTORIAL DESIGNS

- Advantages
  - Multiple interventions tested in one trial (smaller sample size than if two separate trials)
  - Allows examining possibility of interaction effects
  - More participants exposed to potentially beneficial intervention
- Disadvantages
  - More complicated to analyze (must pre-specify whether pooled or four-arm comparison)
  - Very difficult to guarantee no interaction took place, so results can be difficult to interpret
  - Rarely sufficient power to detect interaction effects
  - Power diminished if antagonistic interaction between the interventions

6. FACTORIAL DESIGN EXAMPLE: NEXUS TRIAL

ARTICLES

Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial

Martin Eccles, Nick Steen, Jeremy Grimshaw, Lois Thomas, Paul McNamee, Jennifer Soutter, John Wilsdon, Lloyd Matowe, Gillian Needham, Fiona Gilbert, Senga Bond

Summary

Background Radiological tests are often used by general practitioners (GPs). These tests can be overused and contribute little to clinical management. We aimed to assess two methods of reducing GP requests for radiological tests in accordance with the UK Royal College of Radiologists' guidelines on lumbar spine and knee radiographs.

Methods We assessed audit and feedback, and educational reminder messages in six radiology departments and 244 general practices that they served. The study was a before-and-after, pragmatic, cluster randomised controlled trial with a 2×2 factorial design. A random subset of GP patients' records were examined for concordance with the guidelines. The main outcome measure was number of radiograph requests per 1000 patients per year. Analysis was by intention to treat.

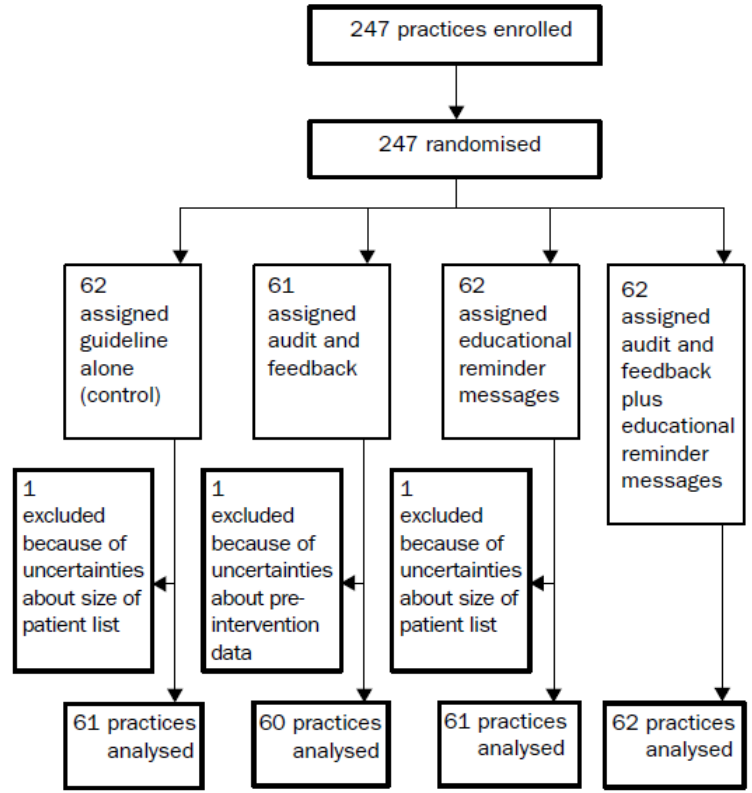
Introduction

General practitioners (GPs) can overuse radiological tests, particularly lumbar spine[1,2] and knee radiographs.[3] Such tests are frequently of little clinical use. Guidelines for use of these investigations are in the UK Royal College of Radiologists' publication *Making the best use of a radiology department*.[4] However, few studies have been done of interventions designed to change GPs' behaviour. Although these studies showed that GPs altered their use of radiological tests, they were badly designed,[5,6] used inappropriate analysis,[7] had short duration of follow-up,[8] or omitted cost considerations.[9] Grol[10] and Lomas[11] have summarised the theory of how to change doctors' behaviour, and Oxman and colleagues[12] have reviewed the effectiveness of interventions. Specific prompts at the time of consultation are a powerful strategy[13] and have been shown to alter GPs' behaviour—eg, when referring patients for infertility investigations[14]—but the effect of the widely-used strategy of audit and

6. FACTORIAL DESIGN EXAMPLE: NEXUS TRIAL

- **Background:** Radiological tests can be overused by GPs and contribute little to clinical management. The NEXUS trial aimed to assess two methods of reducing GP requests for radiological tests in accordance with the UK Royal College of Radiologists' guidelines on lumbar spine and knee radiographs.
- **Interventions**: Audit and feedback, Educational messages attached to X-ray reports sent to GPs
- **Design**: 2x2 factorial design involving 240 family practices served by 6 radiology departments across North East of England and Scotland
- **Outcome**: Number of radiograph requests per 1000 patients per year
- **Results:** Educational messages reduced X-ray requests by 20%, but A&F had no impact.

6. FACTORIAL DESIGN EXAMPLE: NEXUS TRIAL



Trial profile

6. FACTORIAL DESIGN EXAMPLE: NEXUS TRIAL

- "For both types of radiograph, interaction between interventions was not significant—i.e., there was no increased effect of receiving both interventions"

Intervention	Lumbar spine radiographs		
	Before intervention	After intervention	Change
Guideline only (control; n=61)	7·53 (4·1)	6·80 (4·3)	−0·73(2·9)
Audit and feedback (n=60)	7·24 (4·8)	5·97 (4·2)	−1·27 (3·1)
Reminder message (n=61)	7·31 (5·2)	5·14 (3·7)	−2·17 (3·3)
Both interventions (n=62)	8·30 (5·1)	5·23 (3·7)	−3·07 (3·3)

Data are mean (SD).

Table 2: **Radiograph requests per 1000 patients summed across practices for**

AUDIENCE PARTICIPATION – 15 MIN

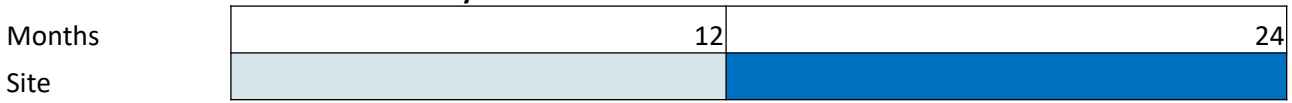
- How should we evaluate the effectiveness of the MND implementation?
  - Consider the 6 different randomized designs with respect to the MND evaluation
  - Discuss possible randomized designs to evaluate the BORN MND intervention

5. NON-RANDOMIZED DESIGNS

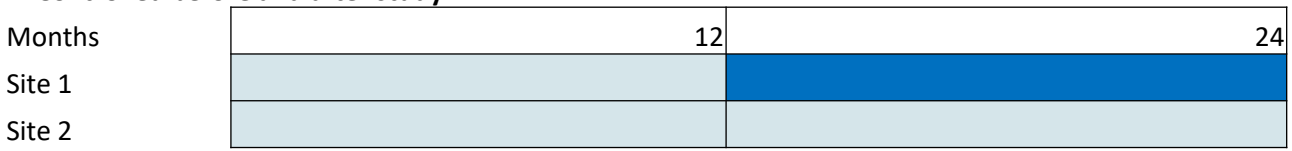
- Major study designs:
  1. Uncontrolled before and after
  2. Controlled before and after
  3. Interrupted time series (ITS)
  4. Controlled interrupted time series
  5. Multiple baseline interrupted time series

NON-RANDOMIZED DESIGNS

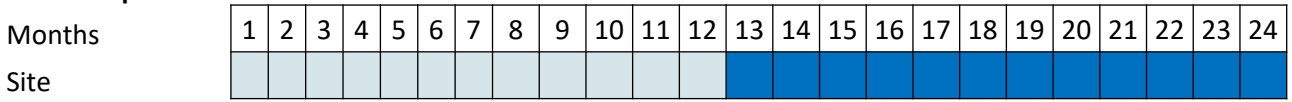
1. Uncontrolled before and after study



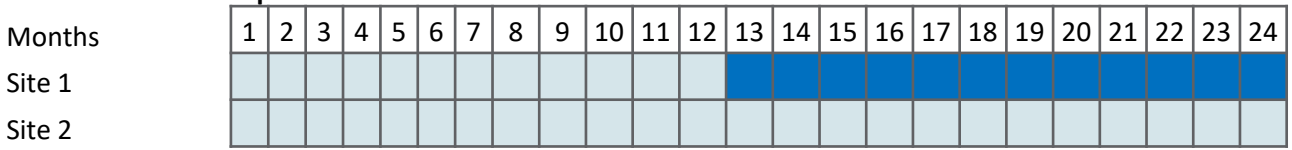
2. Controlled before and after study



3. Interrupted Time Series



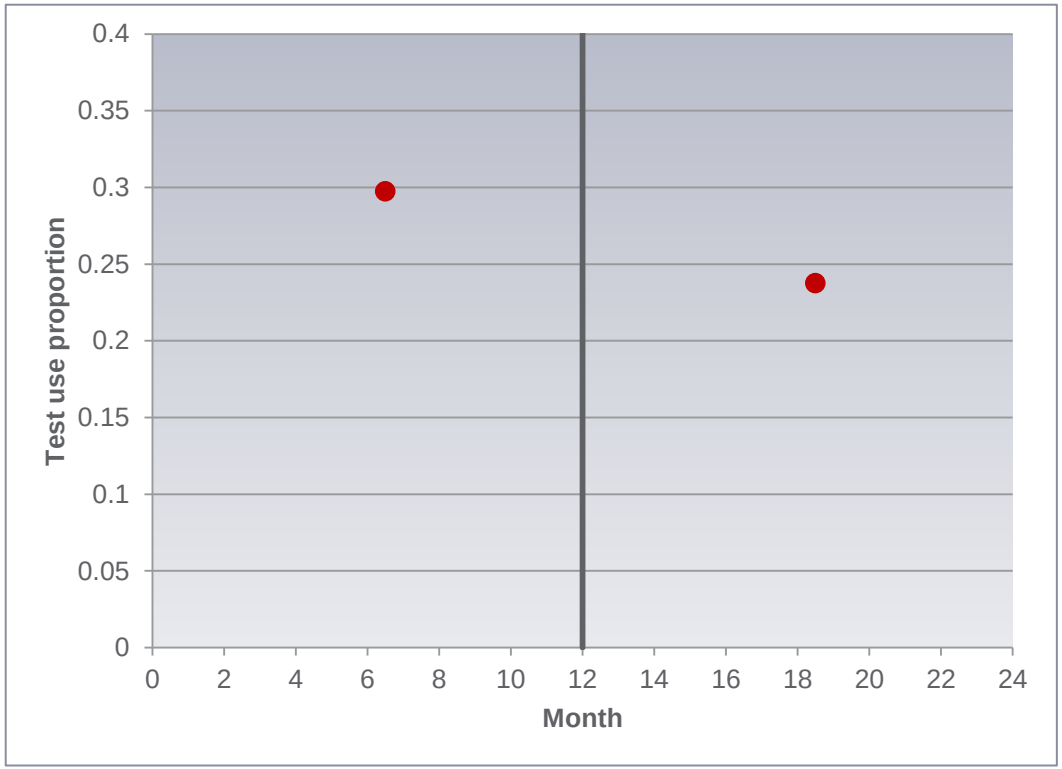
4. Controlled Interrupted Time Series



Try to avoid

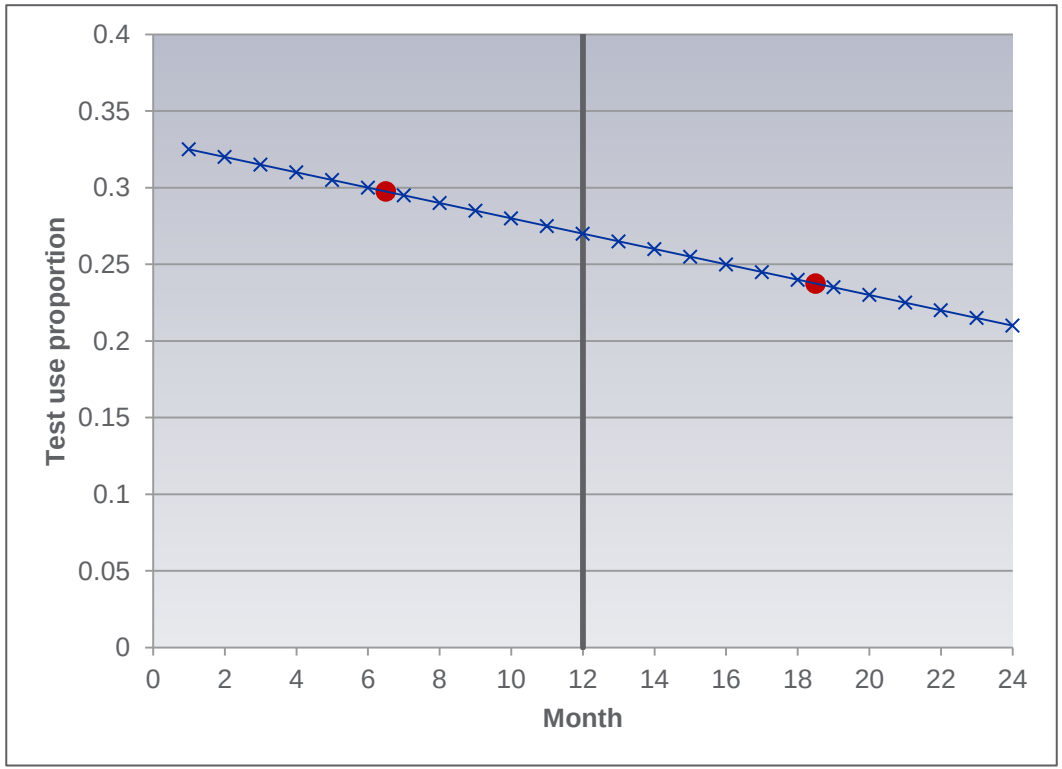
1. UNCONTROLLED BEFORE AND AFTER

- Major threat to validity



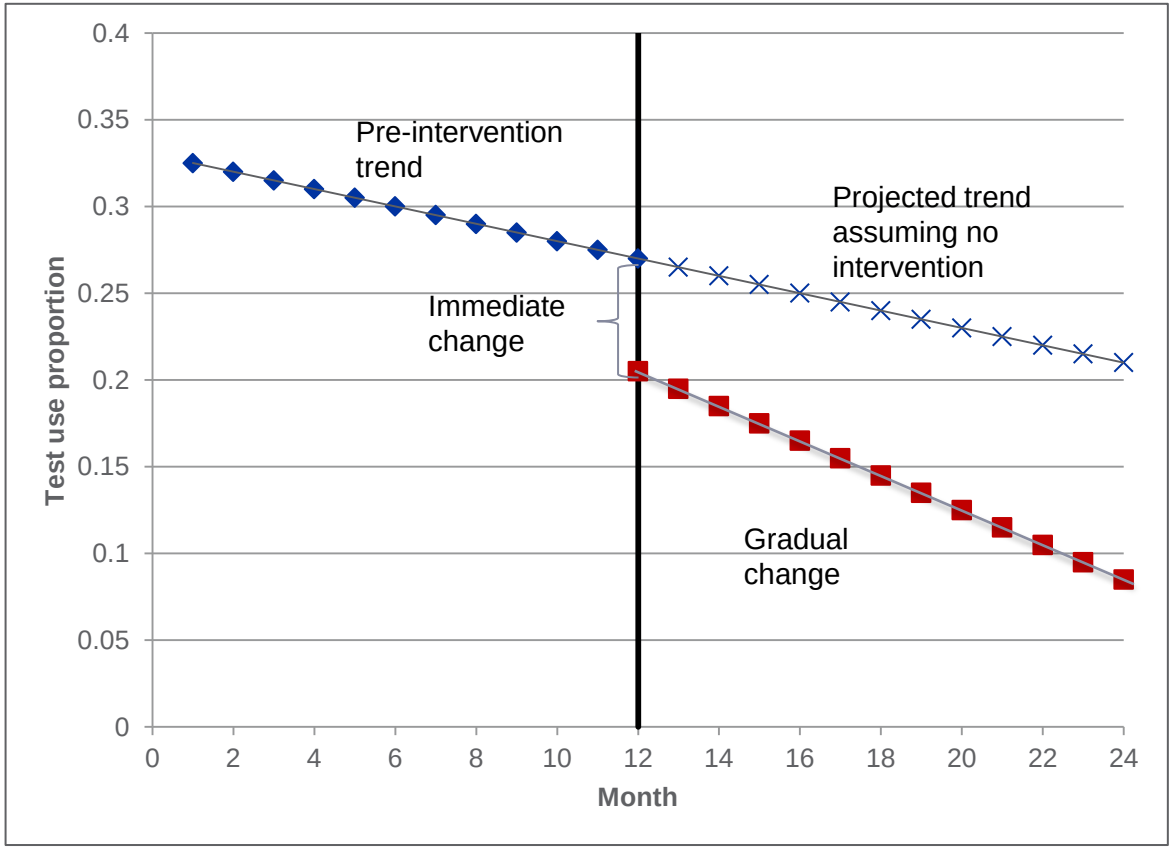
1. UNCONTROLLED BEFORE AND AFTER

- Major threat to validity



Apparent effect completely confounded with the secular trend

3. INTERRUPTED TIME SERIES



- Called "interrupted" time series because we look for an “interruption” in the line at the time of the intervention
- Look for either an immediate change or gradual change
- Can project what outcomes would have been had intervention not been introduced

3. INTERRUPTED TIME SERIES

- Sample size requirements:
  - Single site or multiple sites
  - Need relatively large numbers of observations per measurement (at least 50)
  - Need at least 8-12 measurement intervals pre and post
- Generally more difficult to conduct power calculations

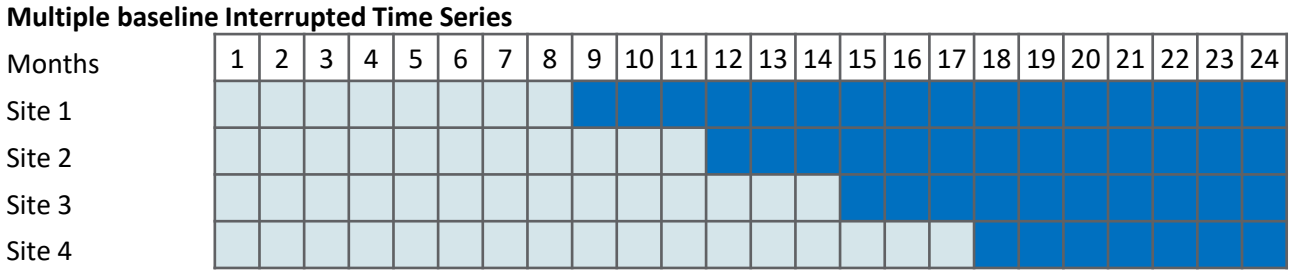
3. INTERRUPTED TIME SERIES

- Advantages:
  - Can be used to evaluate intervention introduced at a single site or at the same time across the population
  - Easy to use with routinely collected data over many time periods
  - Can rule out pre-existing (secular) trends as an alternative explanation
  - Clear graphical presentation of results, easy to explain
  - Only need aggregate data
- Disadvantages:
  - Cannot rule out possibility that another change occurred at the same time as the intervention
  - Long study duration
  - Difficult to interpret when there are few events per time period
  - Difficult to interpret when data collection methods change over time
  - Difficult to separate independent effects of different components of an intervention implemented close together in time

4. CONTROLLED INTERRUPTED TIME SERIES

- Two major threats to validity of interrupted time series:
  - Possibility that another change, occurring at the same time, is an alternative explanation for the observed changes
  - Major shift in the characteristics of the population which coincided with the intervention
- Can be strengthened by adding one or more controls
  - External control: adding an interrupted time-series analysis for a comparison site which did not implement the intervention
  - Internal control: adding an interrupted time series analysis for an outcome not targeted by the intervention
- Compare changes in the control with changes in the intervention series

5. MULTIPLE BASELINE INTERRUPTED TIME SERIES



- Multiple intervention sites with staggered implementation of intervention
- Look for an interruption at a particular time where intervention was introduced, accompanied by absence of an interruption at other sites
- Conduct an ITS analysis in each and pool the results (where possible)
- Looks like a stepped wedge design (but too few sites for stepped wedge)

MULTIPLE BASELINE INTERRUPTED TIME SERIES

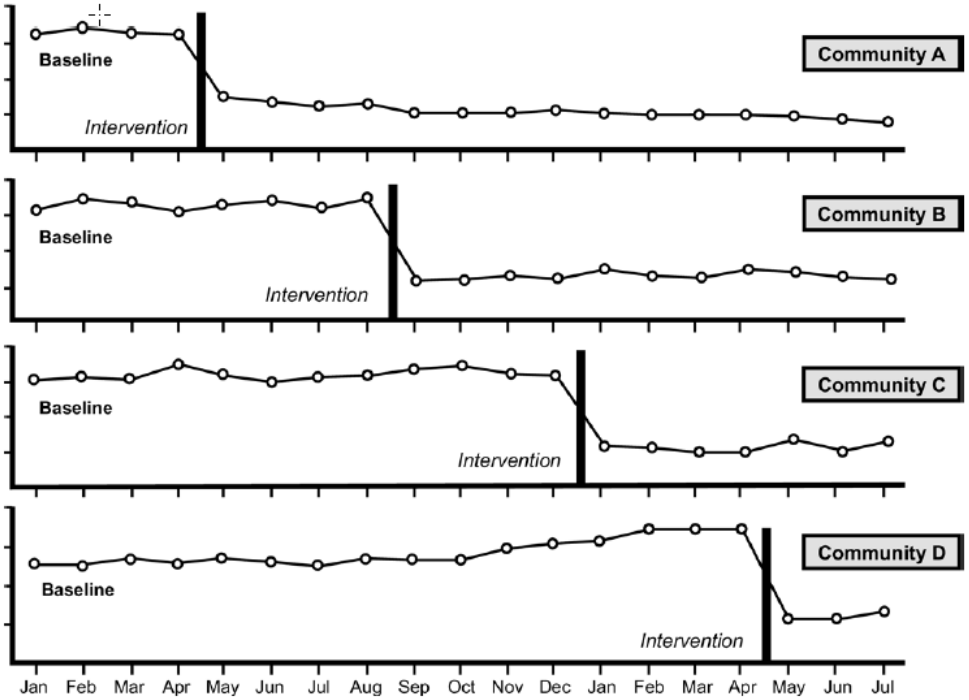


Figure 1. Hypothetical example of a multiple baseline design used to assess behavior change following an intervention in four communities.

5. MULTIPLE BASELINE INTERRUPTED TIME SERIES

- Advantages:
  - Can be used to evaluate intervention introduced at a small number of sites (too few for a randomized design)
  - The greater the number of sites showing a change corresponding to the time at which the intervention was introduced, the more confident one can be that the intervention produced the observed changes (as opposed to some other influences)
- Disadvantages:
  - Can increase the overall study duration
  - Can be difficult to interpret when sites are heterogeneous
  - Works best when different sites operate independently of each other (no contamination)
  - Can be difficult to interpret when interventions are implemented close together in time
  - More difficult to produce a single estimate of intervention effect

AUDIENCE PARTICIPATION – 10 MIN

- How should we evaluate the effectiveness of the MND implementation?
  - Consider the 5 different non-randomized study designs with respect to the MND evaluation
  - Discuss possible designs to evaluate the BORN MND intervention

6. CASE STUDY: BORN-MND

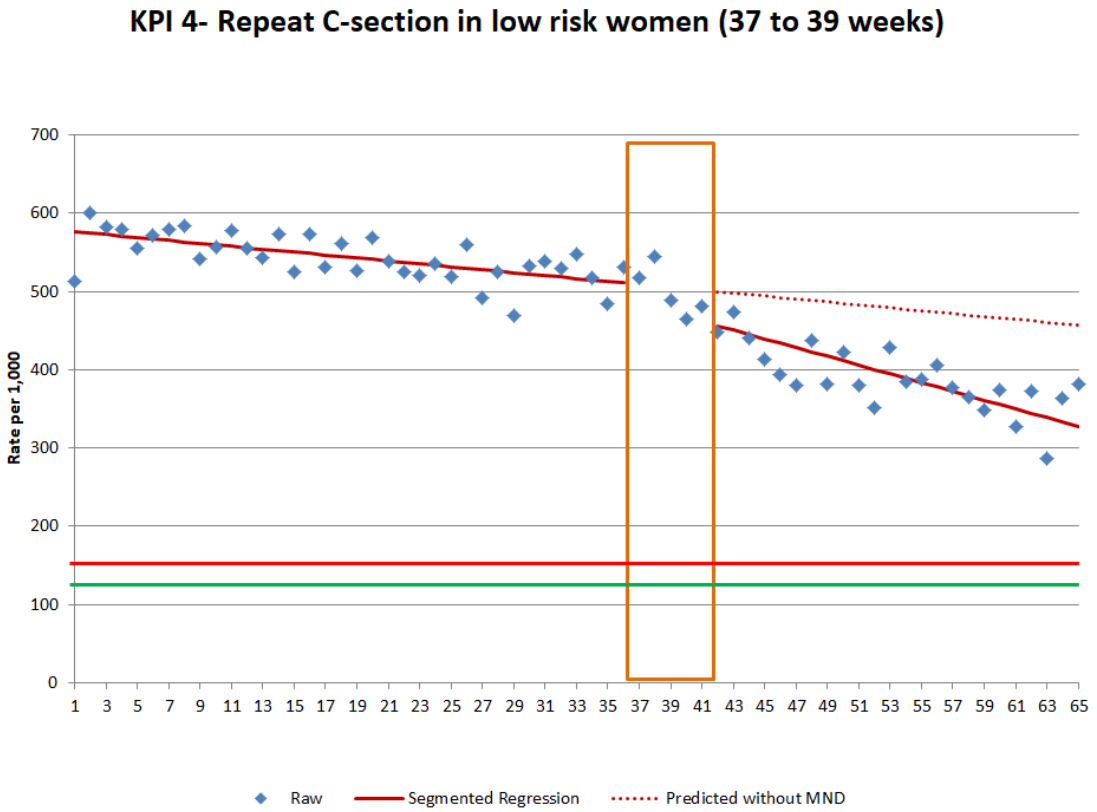
- Randomization not desirable
- MND introduced at all 96 hospitals at the same time
- Selected study design: Controlled interrupted time series analysis
  - Two internal control indictors not targeted by MND
  - Four indicators from external control (British Columbia)
- Study time period
  - 3 years pre-MND implementation and 2 years post-implementation.
  - 5 month implementation period
- BORN registry data 2009-2015 used for all 6 KPIs

RESULTS

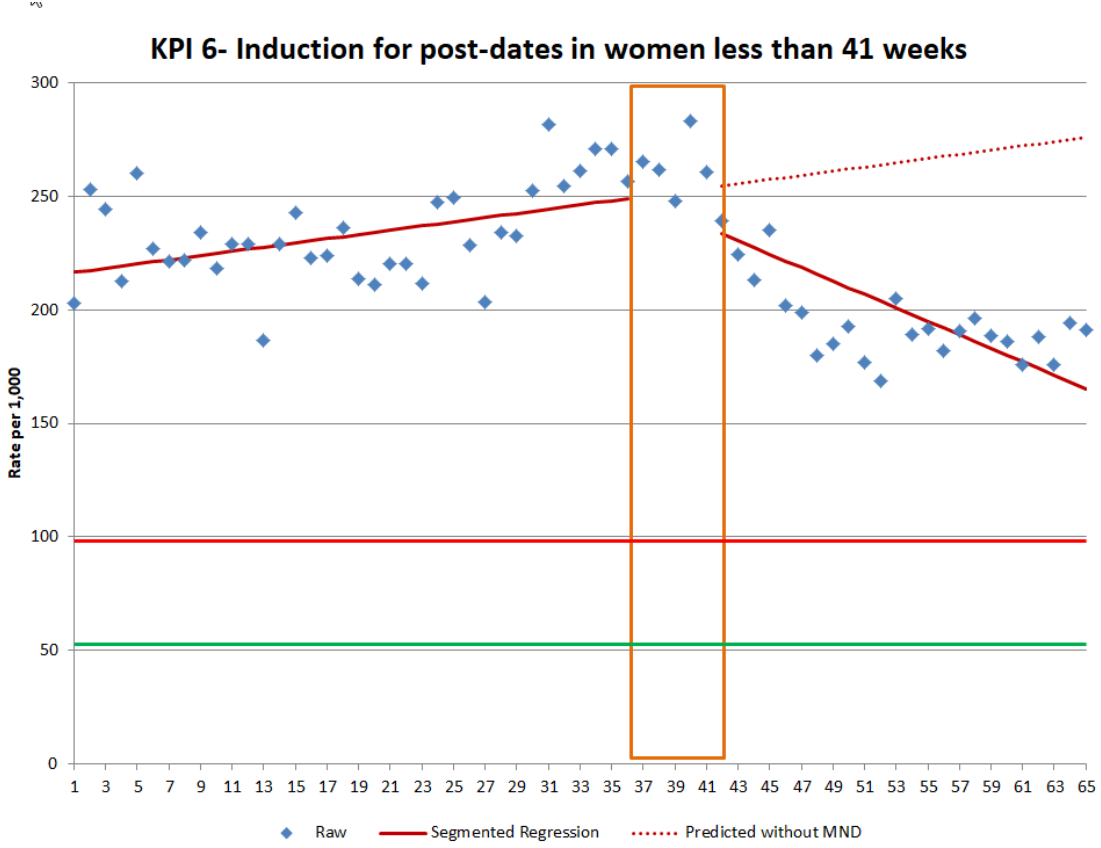
- A statistically significant effect of the MND was found for 4 out of 6 KPIs

- No significant effects were identified for the internal control indicators or in the external control dataset

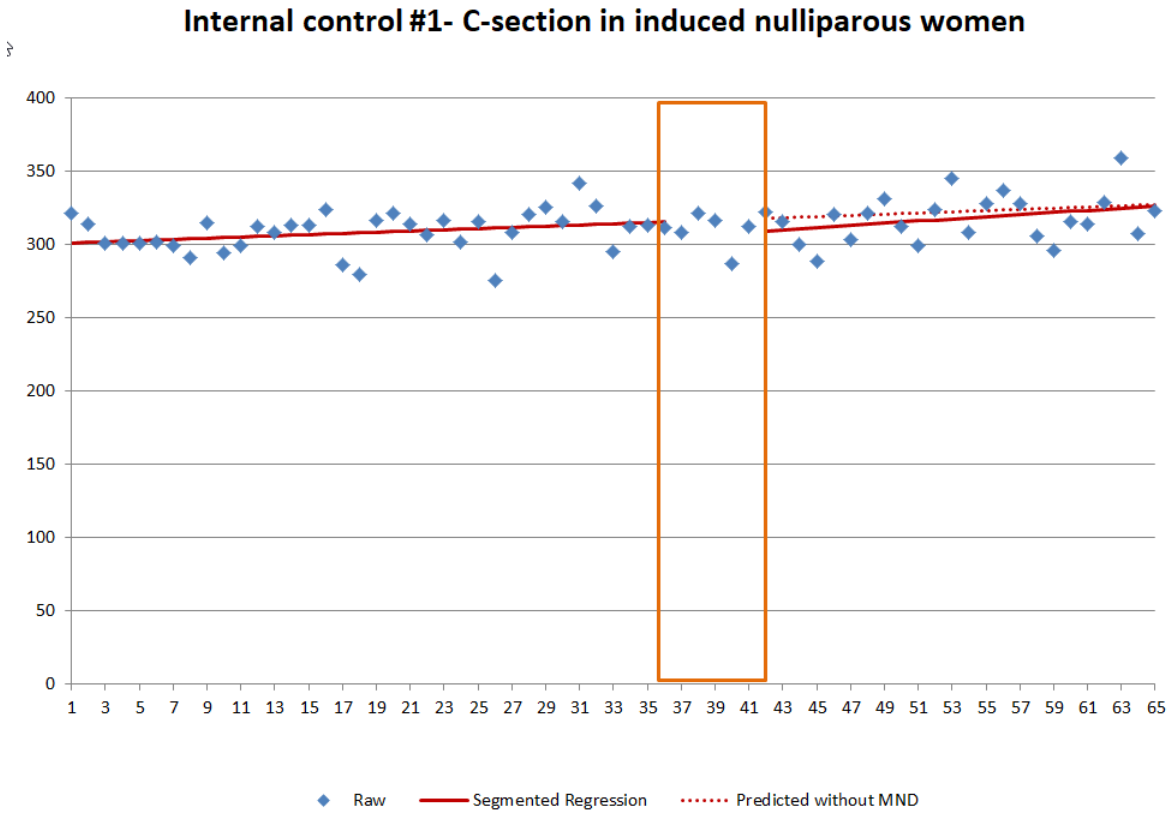
RESULTS: KPI 4



RESULTS: KPI 6



RESULTS: INTERNAL CONTROL



7. FUTURE DIRECTIONS

Ivers et al. Implementation Science 2014, 9:14
http://www.implementationscience.com/content/9/1/14

IMPLEMENTATION SCIENCE

DEBATE Open Access

No more 'business as usual' with audit and feedback interventions: towards an agenda for a reinvigorated intervention

Noah M Ivers[1*], Anne Sales[2], Heather Colquhoun[3], Susan Michie[4], Robbie Foy[5], Jill J Francis[6] and Jeremy M Grimshaw[7]

Abstract

Background: Audit and feedback interventions in healthcare have been found to be effective, but there has been little progress with respect to understanding their mechanisms of action or identifying their key 'active ingredients.'

Discussion: Given the increasing use of audit and feedback to improve quality of care, it is imperative to focus further research on understanding how and when it works best. In this paper, we argue that continuing the 'business as usual' approach to evaluating two-arm trials of audit and feedback interventions against usual care for common problems and settings is unlikely to contribute new generalizable findings. Future audit and feedback trials should incorporate evidence- and theory-based best practices, and address known gaps in the literature.

Summary: We offer an agenda for high-priority research topics for implementation researchers that focuses on reviewing best practices for designing audit and feedback interventions to optimize effectiveness.

Keywords: Audit and feedback, Synthesis, Best practice, Implementation, Optimization

Background

Audit and feedback (A&F) involves providing a recipient with a summary of their performance over a specified period of time and is a common strategy to promote the implementation of evidence-based practices. A&F is used widely in healthcare by a range of stakeholders, including research funders and health system payers, delivery organizations, professional groups and researchers, to monitor and change health professionals' behaviour, both to increase accountability and to improve quality of care. A&F is an improvement over self-assessment [1] or self-monitoring [2] as it can provide objective data regarding discrepancies between current practice and target performance, as well as comparisons of performance to other health professionals. The recognition of suboptimal performance can act as a cue for action, encouraging those who are both motivated and capable to take action to reduce the discrepancy.

The effectiveness of A&F has been evaluated in the third update of a Cochrane review, which included 140 randomized trials of A&F conducted across many clinical conditions and settings around the world. The review found that A&F leads to a median 4.3% absolute improvement (interquartile range 0.5% to 16%) in provider compliance with desired practice [3]. One-quarter of A&F interventions had a relatively large, positive effect on quality of care, while another quarter had a negative or null effect. The challenge of identifying factors that differentiate more and less successful A&F interventions is exacerbated by poor reporting of both intervention components and contextual factors in the literature [4]. Furthermore, most A&F interventions tested in RCTs are designed without explicitly building on previous research or extant theory [5,6]. As a result, there has been little progress with respect to identifying the key ingredients for a successful A&F intervention or understanding the mechanisms of action of effective A&F interventions

IMPLEMENTATION LABORATORIES TO OPTIMISE AUDIT AND FEEDBACK

Reducing research waste with implementation laboratories

The Lancet REWARD (REduce research Waste And Reward Diligence) campaign has encouraged researchers to examine how they work and make efforts to reduce waste and maximise efficiency. Research waste is undermining efforts to improve the effectiveness of health systems. A consistent finding in health services research is inappropriate variations in care and evidence-practice gaps. Implementation science—the study of methods to promote the systematic uptake of clinical research findings and other evidence-based practices into routine practice[1]—can inform health systems on how to reliably improve care and outcomes. However, the potential for implementation science to improve the effectiveness of health systems will not be realised until research waste in the field is systematically addressed.

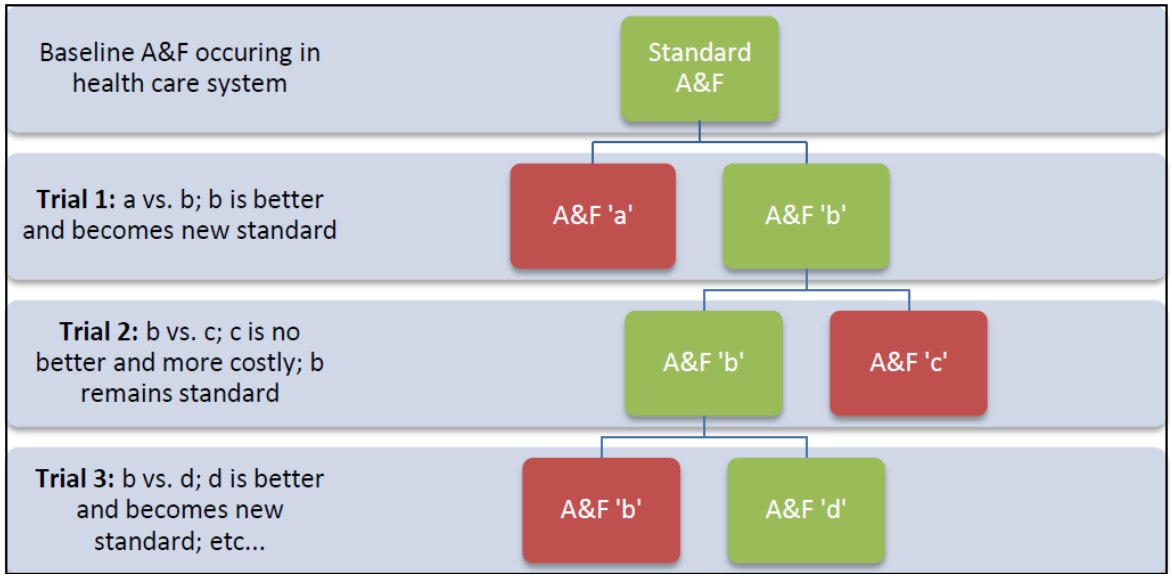
A solid evidence base shows the effectiveness of common implementation strategies—eg, audit and feedback,[2] point of care reminders,[3] educational meetings,[4] and educational outreach[5]—but with substantial unexplained heterogeneity. Yet many current studies that evaluate implementation strategies against control create research waste because they do not build upon the current evidence base or address the key questions to advance the field. For example, for more than a decade we have known that audit and feedback is an effective way to improve care,[2] but researchers continue to undertake trials of audit and feedback versus usual care, testing whether a particular version of audit and feedback can work in a particular setting and for a particular purpose. Such evaluations rarely incorporate relevant theory or best practices[6] in the design and delivery of the intervention and do not address the question of how to optimise the effectiveness of audit and feedback. As a result, there is insufficient evidence on how best to design a new audit and feedback intervention; the same is true for many other implementation strategies.[3–5] Such failures represent substantial waste of scarce implementation research resources and promulgate evidence-practice gaps that incur individual and societal harms.

Health systems have a need for generalisable evidence about how to achieve the greatest possible impact with their quality improvement initiatives.[7] Implementation intervention developers must make many decisions about content, format, and delivery of their intervention; even small modifications in these areas could influence the effectiveness of the intervention.[8] Since the question of whether many common implementation strategies can work has been answered, the time has come for a shift to a comparative-effectiveness model for implementation research.[9] Head-to-head trials that test different ways of designing and delivering implementation strategies are needed to provide the evidence base for health system decision makers. Direct comparisons of implementation interventions will more efficiently move the field forward than the current approach involving cumulating evidence from fairly small trials for indirect analyses in systematic reviews. However, the required sample sizes for such research are difficult to achieve unless the research is embedded within existing, large-scale initiatives.

A promising solution is to develop implementation laboratories that involve close collaboration between health systems delivering implementation strategies at scale and research teams. Implementation laboratories provide an opportunity to kick-start the field by ensuring that scholars meet both applied and scientific goals of understanding what works better and why. Such research can address health systems' priorities and produce generalisable knowledge about factors—context,

For more on the Lancet REWARD campaign see http://www.thelancet.com/campaigns/efficiency

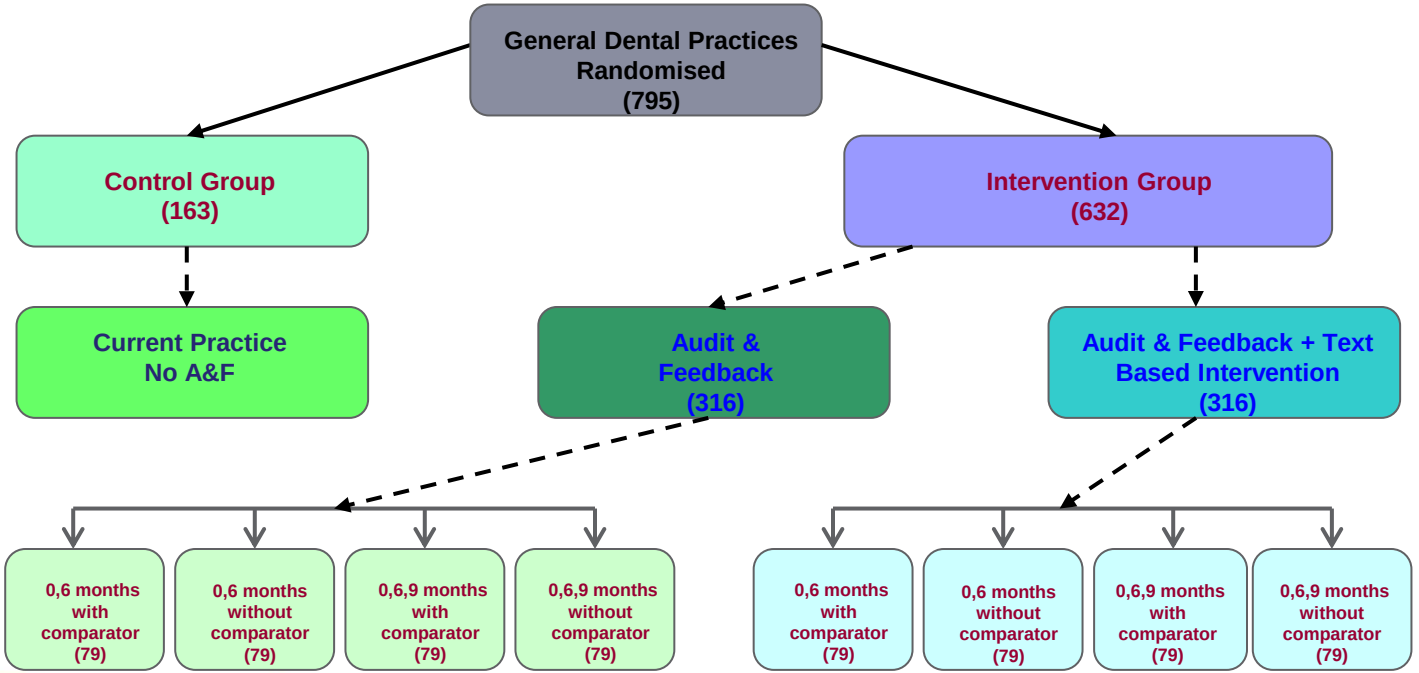
IMPLEMENTATION LABORATORIES TO OPTIMISE AUDIT AND FEEDBACK



IMPLEMENTATION LABORATORIES TO OPTIMISE AUDIT AND FEEDBACK

- Benefits for health system – learning organisation; demonstrable improvements in its quality improvement activities; linkages to academic experts
- Benefits for implementation science – ability to test important (but potentially subtle) variations in audit and feedback that may be important effect modifiers

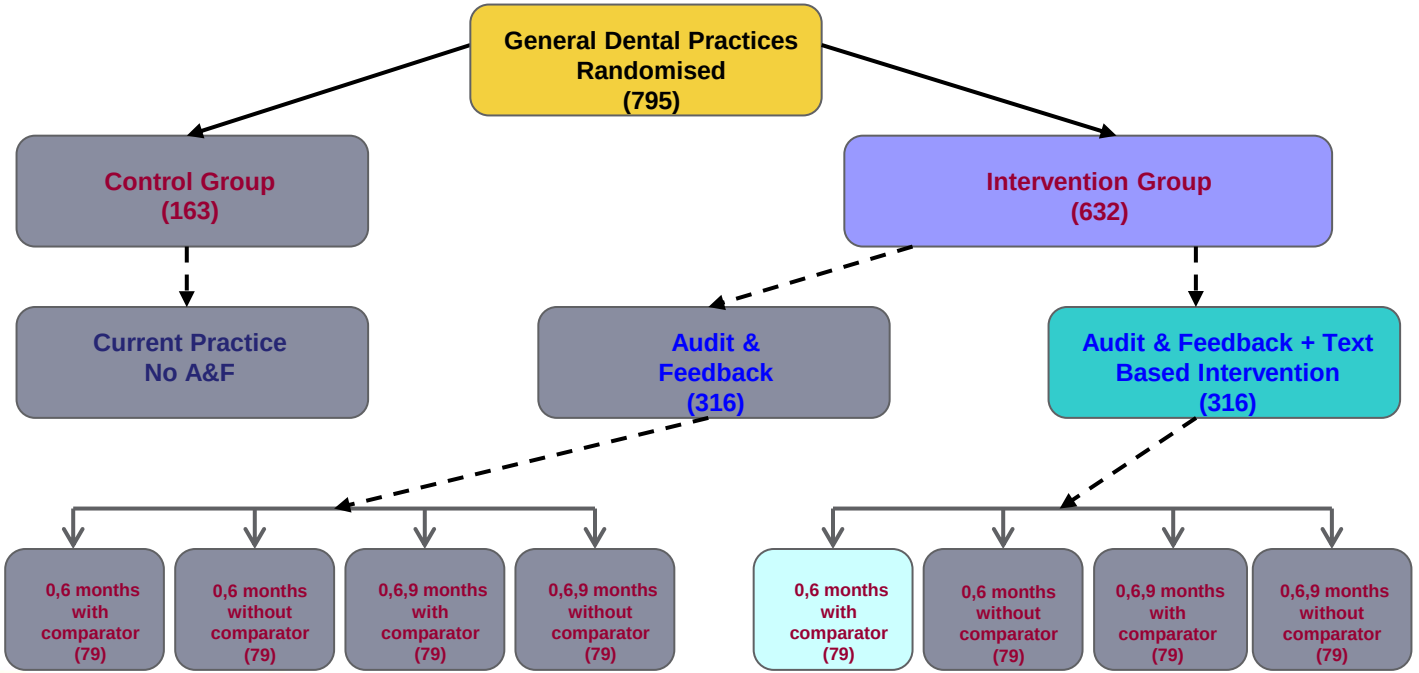
IMPLEMENTATION LABORATORIES TO OPTIMISE AUDIT AND FEEDBACK – RAPID TRIAL



The Ottawa Hospital | L'Hôpital d'Ottawa

Affiliated with • Affilié à uOttawa

IMPLEMENTATION LABORATORIES TO OPTIMISE AUDIT AND FEEDBACK – RAPID TRIAL



The Ottawa Hospital | L'Hôpital d'Ottawa

Affiliated with • Affilié à uOttawa

CONCLUSIONS

- Many possible study designs that have strengths and weaknesses
- Choice of a particular design depends on research question and logistical considerations
- Generally, prefer a cluster randomized design
- Need special expertise to design and analyse appropriately