Assessing the efficacy of mobile phone interventions using randomised controlled trials: issues and their solutions

Dr Emma Beard

2nd Behaviour Change Conference Digital Health and Wellbeing
24th-25th Feb 2016
Background

- Efficacy of mobile phone interventions are generally assessed using exploratory methods (e.g., Pilot studies and proof of concept)

- In an era of evidence based practice calls have been to made to evaluate them with more rigour

- Randomised controlled trials
  - Introduces a number of methodological and statistical analysis issues
Statistical issues

- Missing data – multiple imputation
- Assessing engagement – composite score and survival analysis
- Randomisation failure – regression adjustment and propensity score matching
- Confounding effect of time – random and fixed effect coefficients
- Time-series data – ARIMA modelling
- Underpowered – Bayes Factors
Missing data

- Missing data because of loss to follow-up or non-engagement

- Deletion methods
  - Listwise deletion (complete case analysis)
    - Reduces statistical power
    - Doesn’t use all the information
    - Estimates are biased if data are not Missing Completely at Random (missing value does not depend on X or Y)

- Single imputation methods (mean/mode substitution)
  - Reduces variability
  - Weakens covariance and correlation estimates in the data (ignores relationships between variables)

- Solution: model-based methods such as multiple imputation

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>SES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>18</td>
<td>High</td>
</tr>
<tr>
<td>Male</td>
<td>.</td>
<td>Low</td>
</tr>
<tr>
<td>Female</td>
<td>25</td>
<td>.</td>
</tr>
</tbody>
</table>
Missing data: Multiple imputation

- Data are ‘fill in’ with imputed values using a specified regression model
  - This step is repeated m times, resulting in a separate dataset each time
- Analyses are performed on each dataset
- Results are pooled into one estimate (e.g. Rubin’s Rules)
- Variability is more accurate for each missing value
  - Considers variability due to sampling and variability due to imputation

- A primer on multiple imputation (Schafer, 1999)
- How may imputations are needed? (Graham et al, 2007)
- Potential and pitfalls (Sterne et al, 2009)
- Example: Mobile phone test messages for improving adherence to antiretroviral therapy (ART) (Mbuagbaw et al (2013))
Assessing engagement

• Need to better understand factors influencing engagement (Bailey et al, 2015)
• But how do we assess engagement?
  – Viewing duration
  – Program visits

• Composite measure:
  – Analysis of Cobb et al (2007) Web-based QuitNet smoking cessation program used the product of the number of logins and duration per login in minutes (assumes each measure weighted equally)
  – Statistical weighting methods: multiple correspondence analysis and principle components analysis
    • Handbook on constructing composite indicators (http://www.oecd.org/std/42495745.pdf)
Assessing engagement

- Sometimes, we are interested in how group assignment is associated with engagement and non-engagement
  - t-test (group(control/exp) ~ engagement (yes/no)
- But, sometimes we are interested in how group assignment affects time till non-engagement (duration of engagement)

- Why not linear regression?
  - Group ~ time till non-engagement
    - Survival times are typically positive numbers
    - Cannot handle censoring of observations (Observations of time till an event are incomplete. For example, the study may end before the person has disengaged)

- Survival analysis (e.g., Epton et al, 2013)
  - Traditionally used to assess time to an event eg., death, onset of a disease, length of stay in hospital
  - Cox proportional hazard model (Fisher et al, 1999)
Randomisation failure

- Sometimes randomisation breaks down
  - Subversion bias: manipulation of assignment to groups
    - Schulz et al (1995) badly concealed allocation led to increased effect sizes (cheating?)
  - Technical bias: allocation system breaks down
    - COMET 1 trial (a trial of epidural anaesthetics for women in labour) – program computer fault
  - Attrition bias: attrition differs between groups

- Problem: confounding – participant characteristics which may be associated with the outcome are not evenly distributed
Solution to: randomisation failure

- **Regression adjustment**
  - Regress the outcome on the covariates including an indicator variable for treatment
  - Statistically significant coefficient of treatment indicates a treatment effect
    - Sometimes fails to adequately adjust for differences in observed covariates (Baser, 2007)
    - What about unobserved covariates?

- **Propensity score matching (Rosenbaum & Rubin, 1983)**
  - Propensity score provides a single metric that summarises all the information from explanatory variables such as disease severity and comorbidity
    - It estimates the probability of a subject receiving the intervention of interest given his or her clinical status/sociodemographic characteristics
  - Create a group of treated and control units that have similar characteristics

- Both methods can be combined (Austin, 2011; Austin, 2014)
Derive a propensity score using variables measured at baseline (so they have not been affected by treatment)

Match participants on their propensity score

Analyse these two matched groups

Can reduce sample size due to unmatched participants

Instead of one-to-one matching use one-to-two or one-to-three matching
Confounding effect of time

• In cross-sectional surveys of mobile phone application users, data may be collected at different times

• Need to adjust for time in pooled analyses
  – Responses at time 1 may be different to responses at time 2 as a consequence of external unrelated factors
    • For example, the smoking ban in public places reduced population prevalence rates – introduced confounding if measurements for a mobile phone intervention took place before and immediately proceeding the ban.

• Solutions:
  – Include a fixed effect covariate in our regression model
    • i.e. $Y = \beta_0 + \text{time}_1X + \text{treatment}_2X + e$
  – Include a random effect for time i.e use multi-level modelling
    • If we assume that participants measured at one point in time are more similar (i.e. there is clustering)
    • See Modugno et al (2013) “A multilevel approach for repeated cross-sectional data”
Time-series data

• Mobile phone applications can also provide a proliferation of time-series data
  – Killingsworth and Gibert (2010) iPhone application which tracks various behaviors, cognitions, and affect over time
    • Database contained almost a quarter of a million psychological measurements from individuals in 83 countries.

• Such data generally cannot be analysed with traditional regression based methods due to internal structure e.g. autocorrelation (correlation for a variable with itself at a previous time period) and seasonality
Time-series data

• Need to look at use of more advanced methods from economics and financial arenas

• Dynamic regression models with *Autoregressive integrative moving average* (ARIMA) terms (Hyndman and Athanasopoulos, 2014):
  
  – A regression model can be first fit to the data for explanatory or descriptive modelling
  
  – ARIMA terms can be fit to the residuals in order to account for any remaining autocorrelation
    
    • Prior values *(AR terms)*
    
    • Lingering effects of prior unobserved shocks e.g. white noise *(MA terms)*

  – Can also include seasonal ARIMA terms to account for seasonality
Power of a study

- Chance of finding a significant effect if one exists (usually set to 80%)
  - Higher for smaller standard deviations, larger sample sizes and larger effect sizes

- Before you conduct an RCT you need to calculate the sample size required to detect a clinically important effect

- Sometimes obtaining an adequate sample size can be hard!
Impact of being under powered

- If your study is underpowered you may fail to detect an effect even if one is present (type 2 error)
  - The boy who cried wolf
    - The second error the villagers did (when they didn't believe him) was type 2 error.

- Cannot interpret a non-significant effect:
  - Data are insensitive or evidence for the null hypothesis
Solution: Bayes Factors

• Strength of evidence for some theory versus another
  – If BF > 1 then the data support your theory over the null
  – If Bf<1 then the data support the null over your theory
  – If B=about 1, your experiment was not sensitive

• Jeffreys (1939): BF’s more than 3 or less than 1/3rd provide substantial evidence
  – B>3 substantial support for theory
  – B<1/3 substantial support for the null
To calculate a Bayes Factor

- Online calculator: Zoltan Dienes (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm)

- Baguley and Kaye (2010) provide equivalent R code

- Dienes (2014) “Using Bayes to get the most out of non-significant results”
Acknowledgements

• Thank you for listening

• My salary is funded by the NIHR School for Public Health Research and Cancer Research UK (CRUK)