NIHR Statistics Group Annual Conference "Driving Interprofessional Cooperation"
20-21 June, Sheffield
Breakout Group Summaries

Contents

Title: Are reporting guidelines fit for purpose or are they neither use nor ornament?

Session organiser: Derrick Bennett (on behalf of the Improving Statistical Literacy Working Group)
Number of attendees: ~15
Date and time: 20.6.19, 15.50
Rapporteur: Derrick Bennett

Speaker 1: Michael Schlussel (Senior Statistician, Centre for Statistic in Medicine)

Topic: The **E**nhancing the **QUA**lity and **T**ransparency **O**f Health **R**esearch (EQUATOR) Network

Key points covered

1. The history of the (EQUATOR) network was described.
2. A brief overview of how established guidelines such as CONSORT and STROBE came about
3. A brief overview of guidance on how to develop guidance (e.g. developing a guideline via consensus, producing an exploration and elaboration document)
4. Examples of guidelines specific to statistical issues (e.g. SAMPL, STRESS, DELTA-2, Guidance for Statistical Analysis Plans for RCTs)

Key issues raised

1. Whose job is it to make sure guidelines are adhered to?
2. Several journals endorse guidelines such as CONSORT but do not enforce them.
3. The COMPARE study [Ben Goldacre et al., Trials (2019)] found that when many of the major medical journals (e.g. NEJM, JAMA) were reluctant to publish letters that pointed out poor reporting of published RCTs in their journal.

4. Many people knew of CONSORT but much fewer people knew of the existence of EQUATOR – how could EQUATOR improve its visibility.

Speaker 2: Benjamin Speich (Postdoctoral Fellow, Centre for Statistic in Medicine): EQUATOR Network

Topic: Assessing the impact of reporting guidelines

Key points covered

1. Several studies had assessed the impact on reporting of CONSORT (~80 reports)
2. There were much fewer reports on the impact of other checklists such as STROBE (≤ 10)
3. The burden of completing checklists on authors was mentioned
4. A project was described called ASPIRE that aimed to assess the adherence to the SPIRIT guidelines (a guideline on the content of trial protocols) by studies submitted to research ethics committees.
5. A randomized study was described that aimed to assess the impact on the peer review process of a shortened version of the CONSORT (10 key items) vs usual practice.
6. It was planned to focus on RCTs from the BMJ series and the PLoS series of journals.

Key issues raised

1. Many report of the impact of CONSORT suggest only a modest impact on overall quality
2. How can authors be made to think about the appropriate reporting guideline checklist earlier in the submission process.

Overall Verdict:

Reporting guidelines had improved the overall quality of reporting (particularly of trials) and were useful and informative. But it is the responsibility of journals that endorse them to enforce them. Members of the NIHR Improving Statistical Literacy Subgroup will follow-up with EQUATOR on how to move forward.

Title: Stratification vs Minimisation

Session organiser: Nick Beckley-Hoelscher, Andy Vail
Number of attendees: ~25
Date and time: 20.6.19, 15.50
Rapporteur: Fiona Reid

Key points from presentation:

1. Stratified randomisation provides balance for important prognostic factors, and adjusting for these factors in the analysis generally gives the analysis more 'power'. For continuous outcomes, adjusting improves the precision of the effect estimate. For both binary and survival outcomes, adjusting does not actually improve the precision, but the effect estimate will be closer to the true effect.
2. How many variables should we stratify by? 4 binary stratification factors (for example) means 16 randomisation lists. The more lists we have, the more incomplete blocks we'll have, and therefore less balance over the variables we wish to balance (and in the extreme, if only one patient was allocated within each list this would be equivalent to simple randomisation). Demonstrated by a nice simulation example.
3. Kernan *et al* propose a rule of thumb: For K strata (i.e. list) and block size B, total sample size should be at least 4KB (= an average of 4 complete blocks per stratification list)
4. EMA guidance: "The use of more than 2 or 3 stratification factors is rarely necessary"

5. Instead of stratifying/minimising for centre, consider stratifying for a centre-related feature that may be prognostic, e.g. teaching hospital vs district general

Key items discussed:

1. Several CTUs were represented. Practice varied – some would routinely perform minimisation, some routinely stratification, and for some it was generally a mixture
2. Need a large enough sample in order to stratify for several factors. But if the study is large, then would simple randomisation be just as effective in achieving balance? Not really - the number of patients per strata is the key issue here (e.g. see Kernan's rule of thumb above).
3. Centre/site is often included as a stratification or minimisation factor, without much discussion about whether it is likely to be prognostic. Often included for administrative (not prognostic) reasons. Adjusting for centre as a random effect seemed to be common.
4. Everyone uses a random element in minimisation, usually 80:20. Everyone uses random permuted blocks in stratification. Protocol will not describe block size or minimisation ratio.
5. Minimisation can be too tempting to use, because operationally possible with small samples and many factors, but then run into problems with adjusting for these in the analysis
6. In choosing number of strata for a given sample size, the room seemed to like the 'average 4 complete blocks per list' rule of thumb

---

Title: Variation and repeatability of test results

Session organisers: Sue Mallett and Alice Sitch
Number of attendees: ~12
Date and time: 20.6.19, 15.50
Rapporteur: Emily Robinson

When designing a diagnostic reliability study where at least two professionals (e.g. radiologists) are rating the presence or level of a disease, the following things need to be considered:

1. Interrater/intrarater reliability (consistency) and anonymity of object for review (e.g. image); how to ensure objects are reviewed twice by each rater (ideally without knowing)
2. Pragmatic conditions; how to incorporate the reviewing of study objects into day-to-day clinic to reduce bias of 'test conditions'
3. Choice of patient and clinician populations; how to make sure there is variability in disease status as well as generalisability of subjects and reviewers

Examples mentioned:

1. Poor reliability results from the BUS Study (Accuracy of Bladder ultrasound in the diagnosis of Detrusor Overactivity) [NIHR]
2. Drawbacks of IRR kappa / Intracluster correlation coefficients / Bland-Altman

Tools mentioned:

1. Shinyapps weblink for calculating the reliability coefficient
2. Guidelines for Reporting Reliability and Agreement Studies (GRRAS)
3. Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters (Gwet)
4. Health Measurement Scales: A Practical Guide to their Development and Use (Streiner, Norman and Cairney)

Title: Issues and guidelines in using routinely collected data

Session organisers: Jacqueline Birks
Number of attendees: 24
Date and time: 20.6.19, 15.50
Rapporteur: Emily Robinson

Key Points from the Presentation:

1. Jacqueline Birks (University of Oxford) presented her experiences interpreting the REporting of studies Conducted using Observational Routinely-collected Data (RECORD; www.record-statement.org) guidelines for her CPRD cancer project, with discussion following each slide with members in the room on their experiences.

Key items discussed:

2. We agreed how useful the guidelines were to focus your reporting when writing a manuscript. However, there were some guidelines that some of us couldn't fully adhere to as we didn't have access to the full database, for example in the situation of not having a CPRD site licence.
3. We agreed that reading the full RECORD guidelines would be a good idea when designing a study using routine data, and would recommend that anyone designing one for the first time should take a look.
4. We also discussed issues around sharing code to make our work more transparent and reproducible.

Further action:

1. Adding a resources page to the NIHR statistics website with sources of freely available code lists
2. Organising the next Routine Data meeting
3. Working together on a paper sharing our experiences, with a provisional title of "Big data is not easy data".

Title: Feasibility studies and proof of efficacy

Session organisers: Andy Vail
Number of attendees: 24
Date and time: 20.6.19, 16.30
Rapporteur: Gordon Prescott

Key Points from the Presentation:

1. HTA frequently asks for "proof of efficacy" in an HTA application. How are NIHR statisticians interpreting the requirement and do we need some guidance on what analysis is now suitable for feasibility/pilot studies?
2. Andy gave two examples of rejected applications where a lack of "proof of concept" or "proof of efficacy" was the given reason despite earlier pilot studies. Both could be broken down into four

step mechanistic/causal pathways (or 'logic models') where the evidence for one step was weaker. Andy wondered if this interpretation of lacking 'proof of efficacy' was common at the breakout session.

Key items discussed:

1. All agreed that there is a lack of clarity and that needs to be addressed with the HTA. It would be good to seek clarity from HTA on their interpretation of these phrases and whether they are interchangeable.
2. There was concern that "proof of concept" and "proof of efficacy" were not being used consistently or carefully enough. Some stated that the more important issues were around explaining the causal pathway rather than definitions of terms.
3. There was discussion of the meanings of the terms "proof of concept", "proof of efficacy" and "proof of potential efficacy". There were at least two understandings of the word "efficacy":
    a. evidence of an immediate physiological effect regardless of downstream clinical outcome;
    b. evidence of a downstream effect in carefully selected patients under ideal conditions.

4. There was concern about the language of pharma and basic science being adopted in contexts, such as complex interventions, where it is not necessarily applicable.

Title: Good Statistical Practice Project

Session organisers: Deborah Stocken
Number of attendees: 15-20
Date and time: 20.6.19, 16.30
Rapporteur: Helen Mossopp

Key Points from the Presentation:

1. The presenter gave an introduction and overview of the project; the aim is to develop role-specific GCP training materials for statisticians (primarily aimed at those working in clinical trials) to help them better understand and implement GCP requirements in relation to their roles and responsibilities. The project was initiated and is being run by the UKCRC Registered CTU Stats Operational Group with funding from NIHR (CTU support funding).
2. Training materials have been developed for in-house face-to-face training (to facilitate in-house discussion in relation to any SOPs / processes etc.) and an online tool will also be developed.
3. The presenter piloted a sample of the face-to-face training material focusing on data management and statistical processes. Attendees engaged in discussion and gave feedback on the content which will be useful in updating the material prior to it being released for use.

Key items discussed:

1. There was discussion and feedback on attendee's preferences for methods of delivery (in-house face-to-face, face-to-face external delivery, online) and which method of delivery would work best in their place of work. There was also discussion on whether attendees would prefer this training to be stand alone or a supplement to current GCP training received. Feedback was mixed.
2. All attendees were asked to complete feedback forms which will be disseminated after the session.
3. The training materials are expected to be released for use in Q4 2019.

Further action:

1. Anyone interested in piloting the material (particularly non-trial statisticians) before October 2019 are welcome to contact one of the session facilitators (helen.mossop@newcastle.ac.uk) for more information.

Title: Benefit-Risk Assessment Project

Session organisers: Nikki Totton & Steve Julious
Number of attendees: 15-20
Date and time: 20.6.19, 16.30
Rapporteur: Esther Herbert

Key Points from the Presentation:

1. The Benefit-Risk Assessment to Inform Non-Inferiority and Superiority study design (BRAINS) is a MRC funded methodology project which aims to implement benefit-risk analysis within NIHR/MRC funded projects at the design stage.
2. Benefit-risk analysis is commonly used in pharmaceutical trials where benefits (e.g. clinical efficacy) and risks (e.g. side effects) are more easily defined and often both relate to the patient. However, in publicly funded trials the benefits are sometimes societal, for example cost to the NHS. The project aims to produce guidance for those planning to include benefit-risk assessment in publically funded trials.

Key items discussed:

The groups discussed the current barriers to using such methods.

1. Time to include the required information at the grant application stage – especially if patient input is needed
2. Difficulty in elicitation of preferences from patients
3. Impact on the sample size if using co-primary endpoints or if non-inferiority margin has been changed to reflect the perceived benefit of the intervention.
    a. Perhaps a pipeline/hierarchical approach could be used to avoid having to increase the sample size for multiplicity
4. Access to data to help inform the benefit-risk design
5. Recruitment of patients – difficult to explain the aim of the trial
6. Some benefits are hard to measure. For example, convenience to the patient if investigating the impact of training a patient to perform a task previously done in hospital. Many commonly used QoL measures would not be specific enough to capture this specifically.
7. There was experience in the group of a journal not wanting to publish a benefit-risk outcome as the primary outcome.
8. Everyone needs to be on board

Further action:

1. Going forward this project has a survey (https://scharr.eu.qualtrics.com/jfe/form/SV_1SNrnr7Ivl8pnHD) to gauge the level of current experience with benefit-risk assessment and to guide further work including what should be included in guidance produced.

2. There will be a workshop in September 2019 and interested parties can contact Nikki (n.v.totton@sheffield.ac.uk) or Steven (s.a.julious@sheffield.ac.uk) directly, or indicate interest in at the survey link above.

Title: Machine learning vs classical statistical methods and Routine collected data in clinical trials

Session organisers: Gabriela Czanner, Maria Elstad, Janet Peacock and Catey Bunce
Number of attendees: 35
Date and time: 21.6.19, 11.15
Rapporteur: Elli Bourmpaki

Key Points from the Presentation:

Machine learning vs classical statistical methods

1. A paper on glaucoma patients was presented in which eyes were analysed using statistical methods. The results of this analysis were compared with results from a deep learning method which analysed images from different patients.
2. Both analyses had high accuracy (above 98%)
3. The statistical approach used 50 times less data, and none of this had to be removed due to quality issues. Also, the interpretation of the results was available.
4. The deep learning approach had to remove 15% of the data and the interpretation of results was not available.
5. The need for a collaboration between statisticians and computer scientists was highlighted, to develop a common language and learn from each other's experience.
6. Funding for training to improve communication and negotiation skills is worth applying for statisticians.

Key items discussed:

Routinely collected data in clinical trials

1. There is an increasing interest for using routinely collected data in clinical trials.
2. The most common way to access routinely collected data is via NHS digital, CPRD or HES.
   It is also possible to download data directly for the database at the hospital. However, there is a lot of administrative work needed and it becomes particularly challenging when there are multiple sites involved.

Further action:

3. There is a need for an event dedicated to routinely collected data in clinical trials.

Title: Reproducing code for quality assurance and team working

Session organisers: Carrol Gamble & Sharon Kean
Number of attendees: 25
Date and time: 21.6.19, 11.15
Rapporteur: Louise Linsell

Key Points from the Presentation:

1. The discussion focussed on the Validation of Statistical programming guidelines published on the NIHR website.
2. These guidelines could be used by researchers and statisticians working outside of CTUs
3. There is confusion about the meaning of "validation". Programmers interpret it as the validation of the coding within a program. Statisticians interpret it as the validation of a set of results, i.e. independent verification of a statistic produced from the raw data.
4. There is little, if any, training provided in programming skills and good practice within medical statistics MScs. Statisticians tend to be self-taught programmers and this leads to a wide variation of styles and practices.

Further action:

1. NIHR Statistics group to make the recommendation to include a programming module in MSc courses.
2. Sharon Kean to develop a programming skills webinar and share programming SOP(s)

Title: Quality Improvement and Data Analytics

Session organisers: Abdel Douiri & Brad Manktelow

Number of attendees: 10
Date and time: 21.6.19, 11.15
Rapporteur: Louise Linsell

Key Points from the Presentation:

1. Does not attract key statisticians
2. Many unresolved issues
3. Delivery of current knowledge safely – essential for improvement in health interventions and outcomes

Key items discussed:

1. Missing data
2. Unmatched data in propensity score matching
3. What is the answer we are asking? What question should we be asking?

Further action:

1. Possible half-day meeting on statistical methods in QI
    i. Including non-statisticians
2. Sound out possibility of group (email full membership?)

Title: Data Collection Tools

Session organisers: Clare Lendrem
Number of attendees: 6
Date and time: 21.6.19, 11.15
Rapporteur: Alison Bray

Key Points from the Presentation:

1. Scope of session: research project data as opposed to routinely collected (HES etc).
2. Why now? Rapidly changing data environment, new difficulties being encountered.
3. Overview of relevant regulations: GDPR, GCP, electronic signatures. Definition of *data controller* and *data processor*.
4. Categories of available software: licensed versus free.
5. Options for data location: internet (potential issues with security), NHS network (issues with access), hosted as part of software licence, local, centralised.

Key items discussed:

1. Conducting projects not sufficiently large to qualify for clinical trials unit input and therefore systems, which would be disproportionate, but requiring data collection tools.
2. Research sponsors interpreting regulations differently, leading to inconsistencies which can be problematic for cross-organisation research. Is medical research exempt from requirements of GDPR?
3. Balancing the requirements of data protection with scientific integrity (transparency and reproducibility).
4. Move towards electronic CRFs, no paper source, what are the source data?

Further action:

1. Could the NIHR (statistics group) work towards a consistent interpretation of GDPR? This would require buy in from many different sponsors.
2. Could the NIHR provide a central resource? Challenges would include required staff and infrastructure, defining roles centrally and at sites.
3. A scoping exercise of people at the coal face of research data collection to assess the nature and extent of issues and decide next steps.

Title: Statistical advising

Session organisers: Jamie Sergeant

Number of attendees: 50
Date and time: 21.6.19, 13.15
Rapporteur: Saskia Eddy

Key Points from the Presentation:

1. Victoria Strauss and Catey Bunce shared their experience of statistical advising
    a. Victoria Strauss: involved in several stages of the project from study development & set up to management and dissemination of results.
    b. Catey Bunce is innvolved in a 20 minute or 1-hour session only and if the client requires further support, they will need to pay for this.

Key items discussed:

1. Difficult deciding whether to prepare before an advising session or not.
2. It is important to make it clear at the beginning of the advising session how long the session will be and discuss a few ground rules.
3. Administration support is vital for a successful session.
4. When to start charging for our advice differs from centre to centre.
5. Statistical advising module could be added onto Medical Statistics MScs

---

Title: Studies Within a Trial

Session organisers:

Number of attendees: 25
Date and time: 21.6.19, 13.15
Rapporteur:

Key Points from the Presentation:

1. A background to SWATs was given, the group then reviewed an example statistical analysis plan and discussed key statistical elements relating to SWAT conduct

Key items discussed:

1. Is there any benefit to including a sample size calculation in a SWAT paper. Some felt this would be useful, others felt this might discourage continued work if no evidence of power/effect. Suggested that sample size should be used to indicate how many further trials would be needed for inclusion in a meta analysis to find an effect.
2. What is the difference between a pilot/feasibility study and a SWAT. Some attendees felt that these were closely aligned. The team delivering the session could see this perspective, but suggest that these terms should be separated for avoidance of confusion.

3. Concerns regarding SWAT impact on main trial outcome were raised - for example could providing a pen increase intervention effectiveness. We suggested this is unlikely.

Title: Statistical issues in early phase dose-finding trials

Session organisers: **Simon Bond and Christina Yap**

Number of attendees: 5
Date and time: 21.6.19, 13.15
Rapporteur: Simon Bond

Key Points from the Presentation:

Key advantages of a model-based approach (e.g. CRM) in comparison to a rule-based approach (e.g. 3+3):

1. Statistical rather than deterministic (and therefore requires a statistician!)
2. Decisions based on accumulated information rather than just from the final cohort
3. De-escalation allowed and starting dose does not have to be lowest dose

Key items discussed:

Demonstration of the Shinyapps weblink gave us insight into:

1. Designing a CRM-based trial using simulations of multiple scenarios
2. Being able to provide clinicians with selection probabilities and mean number of subjects dosed
3. Dose Transition Pathways (DTPs); visual schematic diagrams of patient pathways and recommendations made by the CRM model
4. Implementation of a CRM; interactive examples with observed trial data and DTPs for subsequent cohorts