Statistical Analysis of Cumulative Serious Adverse Event Data in Development Safety Update Reports

Harry Southworth
Data Clarity Consulting Ltd
2015/03/23

With many thanks to Dr Brian Davis (HMRA)
Background

  - European regulations require drug developers to submit an annual DSUR for each drug
    - Drugs including biologics
    - With and without marketing approval
    - Commercial and non-commercial sponsors
Contents of DSURs

- Lots of text discussing ongoing clinical trials, cumulative exposure, known, hypothesised, emerging adverse effects, ...
- Case reports of Serious Adverse Events (SAEs)
  - Serious: death, life-threatening, hospitalization, disability, congenital abnormality
- Listings of SAEs since previous DSUR
- **Summary tables of cumulative SAEs (Appendix 6)**
  - AEs typically recorded as free text, then standardized using a Dictionary
  - Often MedDRA ([www.meddra.org](http://www.meddra.org)), but could be COSTART, WHOART or something else
  - **Preferred Terms (PTs), System Organ Classes (SOCs)** and other categories using a dictionary
  - In excess of 17000 MedDRA PTs, in 26 SOCs
  - Many PTs appear to be indistinguishable (e.g. “rash”, “dermatitis”)
  - Quality of data generally low compared to more objective measurements
  - A typical clinical trial will report between about 12 and 1000 different PTs
Problem

- Regulatory authorities currently receiving about 100 DSURs per month
- No systematic, objective approach to reviewing them
- An awful lot of labour-intensive manual review
- Aim to develop an objective, systematic, structured approach to data review

“Statistically guided clinical data review”
Developing a Systematic Approach

- The *only* uniformly structured, complete data is Appendix 6: All SAEs recorded to date
  - Mixes data from multiple studies
  - So, multiple durations, multiple concomitant treatments, multiple disease populations
  - Counts of patients with each preferred term – no patient-level data

- How to systematically analyse?
  - PTs listed alphabetically within SOC
  - Appendix 6 can go on for several pages
    - Hundreds of Pts
    - Need some way of reordering them to draw interesting ones to the attention of reviewers
How not to do it

- Statisticians often worry about chance findings and propose *multiple comparisons* procedures.
- However:

  *Statistical* inference is just one component of the much larger process of *scientific* inference.
Example

- Suppose adjusted or unadjusted p-values are computed for each PT, comparing experimental compound with comparator
  - Draw a cutoff at $p = 0.05$
    - PT with $p = 0.049$ is “Cough”
    - PT with $p = 0.051$ is “Liver failure”
- Would we ignore liver failure and worry about cough?
  - Have drugs like this one been implicated in liver failure before?
  - Does the disease being treated cause people to cough?
- The data can only get us so far
  - The statistician can do little more than rank the PTs to draw attention to the ones that appear to occur with higher frequency on the experimental treatment at this stage
  - Once interesting PTs have been identified, then further analysis might be appropriate:
    - E.g. links to labs, vital signs data
  - Then expert reviewers must combine their experience and knowledge with the data to draw provisional judgements
How to rank

- **The Berry-Berry (2004) model** is designed to model data of this type (i.e. summary AE data)
  - Make use of the MedDRA hierarchy
  - **SOCs assumed exchangeable**
    - PTs are assumed exchangeable within their SOC, given treatment group
    - Allows for borrowing information across SOCs, and between PTs within SOCs
  - Use mixture prior distributions: a spike at 0 for no difference between treatments, mixed with a diffuse Gaussian distribution

- **Popularity of the Berry-Berry approach**
  - According to Google Scholar, cited 74 times (including Southworth & O'Connell)
  - Xia et al modify this approach to work with time-to-event data
    - ... but Appendix 6 doesn't contain that
    - and with highly censored data, it makes little difference
  - Cited as inspiration for DuMouchel
  - Implemented in SAS, S-PLUS
Objections to the Berry-Berry Model

- **Competing hierarchies**
  - PTs within SOCs; PTs can be linked to more than one SOC
  - PTs within Standardized MedDRA Queries
  - Custom, drug-specific hierarchies

- **Exchangeability**
  - PTs have their estimated treatment effects dragged towards each other
    - Ones that are genuinely related to treatment are shrunk towards those that are not
    - ... and vice versa
  - Some AEs will a priori be more likely than others (but how to put individual priors on hundreds of AEs?)
  - DuMouchel:  
    
    *the different safety issues should be medically related, so that it is plausible that the different issues have related mechanisms of causation or are different expressions of a broad syndrome*

- **Example: Vascular Disorders**
  - Wound haemorrhage, varicose vein, hypotension, circulatory collapse, ...
Further objections

- Choice of prior
  - Point mass at no difference between treatments
  - Some SAEs might be anticipated
  - Could be knowledge to the contrary from elsewhere
  - In general, possibly more appropriate to doubt that any 2 treatments are identical; the question is whether the difference is big enough to detect or care about
What to compare Berry-Berry with

- Something simple that avoids all those objections
- Need to deal with the case in which 0 are observed in one of the treatment groups
  - Add $\frac{1}{2}$ an event and $\frac{1}{2}$ a non-event to each PT
    - $p_x = \frac{x + \frac{1}{2}}{N_x + 1}$, $p_y = \frac{y + \frac{1}{2}}{N_y + 1}$
    - $rr = p_x / p_y$
  - The Jeffreys prior, but can be justified from several other viewpoints
- So, just simple estimates treating every PT as being independent of all others
What to report

• From both the Berry-Berry model and the simple approach:
  – PTs ranked by relative risk, lower 90% confidence limit of relative risk
  – Why relative risk? Why not risk difference, odds ratio?
    • Not much reason. Possibly preferred by clinical reviewers and the purpose is
      “Statistically guided clinical data review”
      and relative risk will be fine for that
  – Why 90% intervals?
    • Attempt to avoid their use as surrogates for 5% level hypothesis tests
      (Sterne & Davey-Smith, Cox, 2001)
    • (I'm told this really is a problem!)
Evaluation

• Experience with old marketed drugs revealed no real role for a new approach
  – Many known AEs expected, amount of new data since previous DSUR small, non-comparative

• Simulated data not helpful because no way to make clinical (i.e. not statistical) judgements

• Appendix 6 data provided from 6 DSURs
  – Blinded to the treatments
  – Provided as non-uniform Excel files
  – First step is to turn the data into something usable via Python

• Standard reports returned to expert reviewer for evaluation
Results

- The Berry-Berry model is dominated by its priors
  - Sample sizes 2300, 1100, 349 PTs

<table>
<thead>
<tr>
<th>Condition</th>
<th>5%</th>
<th>25%</th>
<th>RR</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aplastic anaemia: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.91</td>
</tr>
<tr>
<td>Coagulopathy: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Febrile neutropenia: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.99</td>
</tr>
<tr>
<td>Haemolysis: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Haemolytic anaemia: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.39</td>
</tr>
<tr>
<td>Leukopenia: {0.2%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>4.41</td>
</tr>
<tr>
<td>Pancytopenia: {0.2%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.92</td>
<td>4.91</td>
</tr>
<tr>
<td>Splenic infarction: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.40</td>
</tr>
<tr>
<td>Acute myocardial infarction: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>2.02</td>
</tr>
<tr>
<td>Arrhythmia: {0.2%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>3.51</td>
</tr>
<tr>
<td>Atrial flutter: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.42</td>
</tr>
<tr>
<td>Bradycardia: {0.3%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.88</td>
<td>4.45</td>
</tr>
<tr>
<td>Cardiac failure acute: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cardiac tamponade: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.01</td>
</tr>
<tr>
<td>Cardiogenic shock: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Cardiopulmonary failure: {0.1%, 0%}</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Comments

• The point-mass mixture prior dominates many PTs
  – Many interval estimates of relative risk 1 with 90% posterior interval (1, 1)
  – This was a surprise – my main concern had been non-exchangeability

• Turns out, I'm in good company:
  – Evans, Prieto-Merino, Spieglehalter, Whittaker and Smeeth (2008) had also found undesirable behaviour due to priors:
    “The spiked priors can produce bimodal posteriors that are difficult to explain clinically”

The simple approach

- Identified many of the safety concerns identified by the sponsors
- Identified several new leads for reviewers to look into
- Judged to provide an objective, systematic and efficient approach to review of Appendix 6
But why not just use vague priors?

- I still worry about the exchangeability assumption
  - If there is modest evidence a specific event is related to treatment, and if in the light of the data it is considered a plausible relationship, why dilute it?
  - The 'correct' groupings are not known a priori, but that does not imply then don't exist or cannot be figured out later
    - “The reason I didn't prespecify the hypothesis is that I hadn't thought of it then. Now that I have, it is most interesting.” [A sentiment borrowed from Edwards]
- The simple approach is easy to implement, hopefully assisting take-up
- No matter how sophisticated the statistical methodology is, the quality of the data is low
- The purpose is “statistically guided clinical data review” and the output of any statistical ranking must be judged in the light of other available information and experience
References

- S. M. Berry and D. A. Berry, Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model, Biometrics, 60, 418 – 426, 2004
- D. R. Cox, Another comment on the role of statistical methods, British Medical Journal, 322, 231, 2001