Extreme value modelling:
A novel approach to the analysis of clinical trial safety data

Work resulting in the 2012 award for statistical excellence in the pharmaceutical industry

Harry Southworth
AstraZeneca
2013-11-25

Acknowledgments: Janet E. Heffernan, Yiannis Papastathopoulos, Jonathan Tawn

AstraZeneca

THE TIMES

SATURDAY

Max 15C, min 3C | November 21 2009 | timesonline.co.uk | Newspaper of the Year | No 69799 | £1.50

Once every 1,000 years rain falls like this

Hundreds of homes evacuated after floods

Steve Bird, Lindsay McIntosh

The full and devastating impact of England's worst recorded day of rain was still emerging last night as tributes were paid to a policeman swept away by floodwaters while trying to save others.

PC Bill Barker was helping motorists stranded on a bridge over the Derwent in the Cumbrian town of Workington when it collapsed. His body was discovered hours later on a nearby beach.

homes and businesses were evacuated, many of them ruined by floodwater and mud.

Jerry Graham, Cumbria's Assistant Chief Constable, said that PC Barker and a colleague had gone on to the bridge to help drivers who were trying to cross it. He said: "It was obvious they were going to put themselves in danger so PC Barker went to try and protect them. The bridge gave way just due to the volume of water and PC Barker went into the water and was swept away."

Overview

- Context
- Extreme value modelling
- Evaluation
- Implementation
- Selling to the business
- Impact

Drug development and safety data

- Loosely, drug development takes place in 3 phases
  - Phase 1 – small numbers of healthy volunteers, checking tolerability
  - Phase 2 – more patients, finding right dose
  - Phase 3 – many patients, confirming efficacy
- Trials designed to characterize efficacy, but most data relate to safety
  - Adverse events – did the patient notice any side effects?
  - Lab data – measurements on various chemicals in the blood
  - Vital signs – heart rate, blood pressure
  - Medical history, concomitant medications, ECGs, etc.

AstraZeneca

The problem

- The pharmaceutical industry is in crisis
  - The cost of drug development has been rising steadily for years
  - The difficulty of getting a drug to market has been rising steadily for years
- The second most common cause of late phase failure is safety
- The most common cause of market withdrawal is safety

AstraZeneca

Why aren't safety issues identified sooner

- Difficult to address in a formal hypothesis testing framework
  - Safety questions are generally not known in advance
  - There are lots of them
  - Concerns over multiple comparisons
- The data are messy
  - Binomial or time-to-event data, often sparse
  - Continuous data often highly skewed and subject to outliers
  - Data collected at several points in time
- Usually, only the outliers are of interest
  - Methods concerned with characterizing central tendency can imply misleading conclusions
  - Outliers are, by definition, rare

Outliers and ALT

- The rest of the presentation focusses on outliers and on ALT (alanine aminotranferase).
  - Large values of ALT suggest potential liver injury
  - Potential for liver injury has been the most frequent reason for safety related withdrawl of drugs from the market [1]
    - *ticrynafen*, *benoxaprofen*, *bromfenac*, *troglitazone*, *nefazodone*, *ximelagatran*...
- According to published guidance (CTC [4]):

	Grade	**Severity**
$ULN \leq ALT < 2.5 \times ULN$	1	Mild
$2.5 \times ULN \leq ALT < 5 \times ULN$	2	Moderate
$5 \times ULN \leq ALT < 20 \times ULN$	3	Severe
$ALT > 20 \times ULN$	4	Life threatening

  - ULN can be thought of as the units of measurement for ALT

Example: troglitazone

Troglitazone (for treatment of diabetes). FDA review states

> *Mean [ALT] levels fell in patients receiving troglitazone in phase 3 trials... It was also stated that 2.2% of patients in phase 3 trials had an [ALT] level exceeding* $3 \times$ *ULN... What was not appreciated by [FDA] was that many of the patients classified as ALT* $> 3 \times$ *ULN actually had ALT values that were VERY much higher than* $3 \times$ *ULN... 23 patients had treatment-emergent ALT values over* $3 \times$ *ULN... In 14 of these 23 patients, the ALT value exceeded* $8 \times$ *ULN... and in 5/23 patients the ALT value exceeded* $30 \times$ *ULN.*

The drug was withdrawn from the market after reports of liver failure and death

AstraZeneca

A short aside

- In fact, the liver data is at least 4-dimensional
  - Need to look at least at AST, bilirubin and alkaline phosphatase as well
- But ALT contains most of the information
  - ... and keeps this presentation fairly simple
- See Southworth & Heffernan [9] for the multivariate version

Extreme value modelling

- A mature branch of statistics
  - Key publication by R. A. Fisher and L. H. C. Tippett, 1928
  - First full length textbook by E. Gumbel, 1958
  - In the context of predicting the frequency and severity of floods, Gumbel wrote:

    "At present the statistical nature of these problems has been realized and the empirical procedures are slowly being replaced by methods derived from the theory of extreme values"

- These days, the methods are used in many areas of application:
  - Meteorology, insurance, finance, geology, metallurgy, breaking rates of fibers, network traffic, ...

AstraZeneca

The generalized Pareto distribution

- Asymptotically, a threshold exists above which the data are well approximated by a generalized Pareto distribution (GPD)
  - $F_{>u}(x) = 1 - \left\{1 + \xi\left(\frac{x-u}{\sigma}\right)\right\}^{-1/\xi}$ for $x > u$
  - ξ: the shape parameter
    - $\xi < 0$: short tailed distribution
    - $\xi \geq 0$: heavy tailed distribution
- The scale and shape parameters don't have a straightforward physical meaning
  - Therefore present results in terms of predictions from the model
    - Predicted probabilities of exceeding certain thresholds
    - Predicted extreme quantiles of the distribution – **return levels**
- Can allow σ and/or ξ to depend on covariates

AstraZeneca

The need for evaluation

The methods are in widespread usage, but:

- Apparently unused in the context of clinical lab data
- Asymptotic justification for GPD, but clinical trials are 'small'
- The ideas feel quite alien to most statisticians and clinicians alike
  - Throw away at least half of the data
  - Extrapolating outside the range of the data
- Various issues relating to clinical trials don't arise in other applications
  - See Southworth & Heffernan [8] for a more complete account
- If it works, will need concrete examples to sell the idea to colleagues

Suggests some evaluation should be performed

AstraZeneca

Evaluation

Approach:

Use existing clinical trial data to fit GPD models and predict frequency and magnitude of outliers in established drugs whose impact on ALT is known.

- Identified 3 old clinical trials with appropriate data, including some data from patients receiving placebo
- Began modelling using the *ismev* [3] package for R [7]
  - (Others exist)
- Quickly ran into trouble

Issues with extreme value modelling of clinical trial data

- No reason to suppose quadratic approximations are good
  - Can use profile likelihood for inference [2]
  - ... but profile likelihood gets awkward when there are multiple parameters in the model
- MLEs can fail to converge (especially in the sample sizes we encounter)
  - Fewer convergence issues if reparameterize in terms of $\phi = \log \sigma$
- No extreme value methods exist(ed) for longitudinal data
- The appropriate multivariate method [5] had no implementation
  - and had a known glitch [6]

AstraZeneca

Approaches

Write new software: `texmex` [10]

- Use penalized likelihood, Bayesian or bootstrap approaches to inference
  - What choice of penalty or prior?
  - Empirical bootstrap badly behaved, need parametric bootstrap
  - Use R's usual formula interface for model specification
- Use on-treatment maxima and ignore longitudinal structure
  - Or use data from a single on-treatment clinic visit
- Implement the Heffernan-Tawn model [5]
  - Made contact with the authors and have worked with them ever since

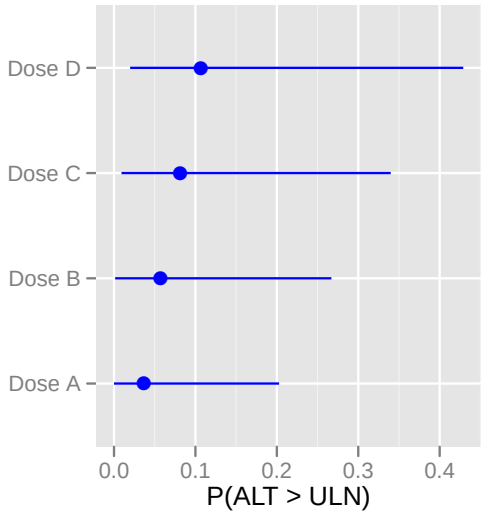
Various other issues eventually addressed as part of a PhD

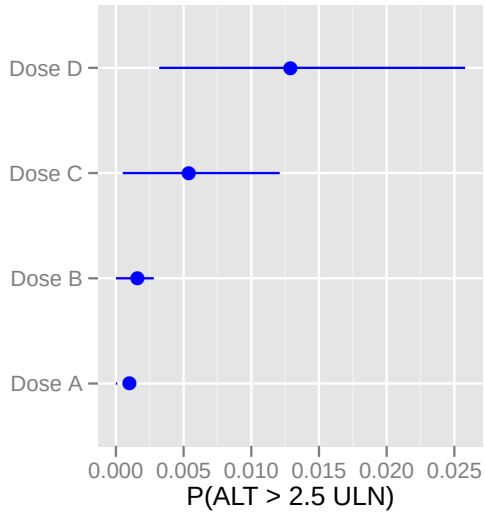
- ... and others remain to be addressed

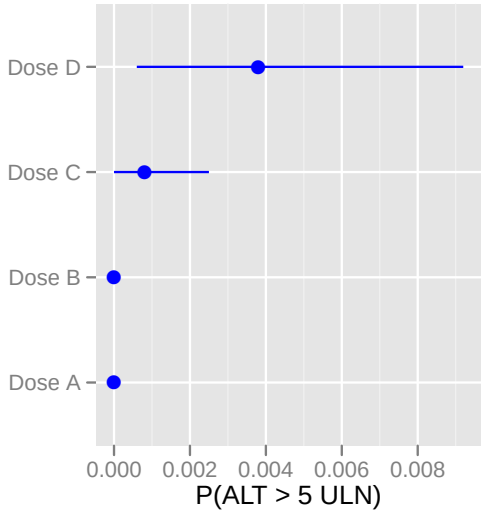
AstraZeneca

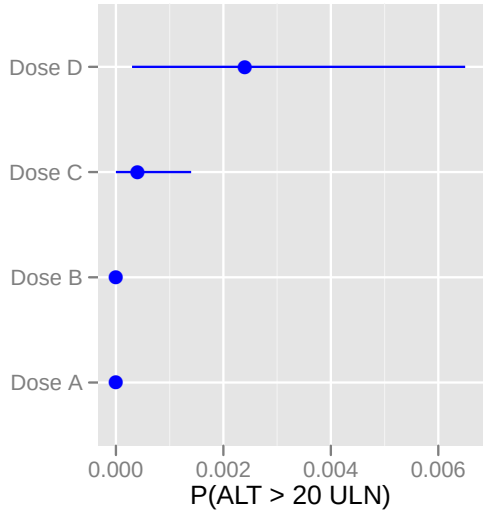
On with the evaluation

- For 2 of the clinical trials used for evaluation, predictions for ALT were in line with what is known in practice
- For the other, I found something I hadn't expected
  - There were 4 doses of the drug, approximately 160 patients per dose
  - Models predicted approximately 1 in 400 patients taking the highest dose would have an $ALT > 20 \times ULN$
  - Knowing the drug, I didn't believe it
    - Remember CTC classification:
      $ALT > 20 \times ULN =$ "life threatening"









AstraZeneca

Literature search

- Spent a week checking my code
- Gave up and did a literature search
  - Found published information that approximately 1 in 400 patients had a severe liver related adverse event when taking the highest dose in clincial trials
  - No cases of liver related events on the lowest dose
- From a study of approximately 160 patients per group, extreme value modelling had correctly predicted the rate of serious liver related events

Selling the idea within AstraZeneca

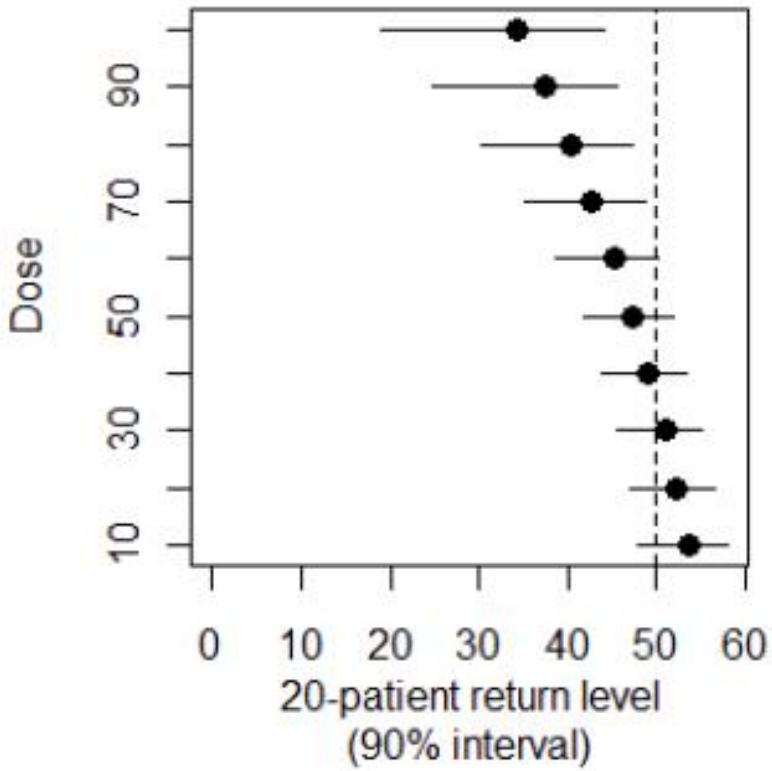
- Audience is advisory panel on liver tox: largely medics
- Need to anticipate and address the concerns of the audience
  - Extreme value modelling feels uncomformable
    - It's old, it has pedigree, it has been used in many important applications for a long time
    - But it's just a limit theorem and no one worries that means aren't Gaussian
  - Emphasize prediction, not hypothesis testing
  - Need to explain what it can offer that they don't already have
    - Can predict the 99^{th} percentile of ALT from a few dozen observations
    - Can predict $P(ALT > 3 \times ULN)$ even if we have seen no cases
  - ... and the 1 in 400 patients example won the day

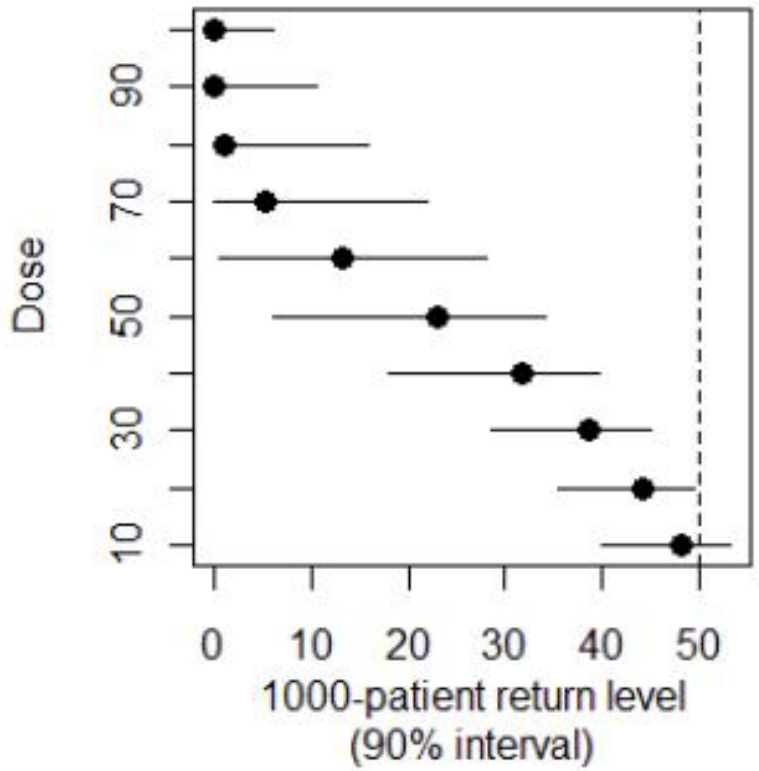
AstraZeneca

Next steps

- Got to present to the lead medic and lead statistician in AstraZeneca
  - They got it!
  - Asked for analysis of phase 2 ximelagatran data
- Started to use for real with a couple of examples
  - Mostly liver-related stuff (ALT)
  - ... but then a phase 2 study threw up a few big changes in left ventricular ejection fraction (LVEF), and someone involved had heard of my work with extremes
    - LVEF is the proportion of blood pumped out by the heart
    - If it gets low, that's bad - *heart failure*
    - So we're interested in extreme small values, not big values, so multiply by -1 before modelling

AstraZeneca





AstraZeneca

Further roll-out

- AstraZeneca's internal governance and advisory bodies commonly request extreme value modelling to be performed
- The most common examples include
  - ALT, LVEF, creatinine, neutrophils
- Most of the modelling is performed by a specialized group of statisticians
  - Avoids having to train dozens of statisticians in a niche area of modelling they might never use
  - Allows the specialized group to gain experience quickly
  - Enables faster execution
  - Makes clear who to go to to ask for modelling to be done

AstraZeneca

Closing remarks:
Extreme value modelling of clinical trial safety data

- Experience to date suggests
  - Extreme value modelling really can predict toxicities from early phase data
  - Approximately 50% of the time, outliers that appear worrying turn out to be consistent with no treatment effect
    - "Proceed with caution" rather than "clean bill of health"
- Perhaps it is the case that

"the statistical nature of these problems has been realized and the empirical procedures [will slowly be] replaced by methods derived from the theory of extreme values"

AstraZeneca

And finally

- The “1000 years” remark was made by then Environment Secretary, Hilary Benn
- According to Jan Heffernan’s analysis of rainfall data from 2 weather stations in the region, once every 1000 is a plausible estimate

References I

U.S. Food and Drugs Administration.

Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation.
2008.

S. Coles.

An Introduction to Statistical Modelling of Extreme Values.
Springer, 2001.

S. Coles, J. E. Heffernan, and by A. G. Stephenson.

ismev: An Introduction to Statistical Modeling of Extreme Values, 2012.
R package version 1.39.

DCTD, NCI, NIH, and DHHS.

Cancer Therapy Evaluation Program, Common Terminology Critera for Adverse Events, Version 3.0.
2003.

J. E. Heffernan and J. Tawn.

A conditional approach for multivariate extreme values.
Journal of the Royal Statistical Society Series B, 56:497 – 546, 2004.

C. Keef, I. Papastathopoulos, and J. A. Tawn.

Estimation of the conditional distribution of a vector variable given that one of its components is large: additional constraints for the heffernan and tawn model.
Journal of Multivariate Analysis, 115:396 – 404, 2013.

AstraZeneca

References II

R Core Team.

R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2013.

H. Southworth and J. E. Heffernan.

Extreme value modelling of laboratory safety data from clinical studies.
Pharmaceutical Statistics, 11:361 – 366, 2012.

H. Southworth and J. E. Heffernan.

Multivariate extreme value modelling of laboratory safety data from clinical studies.
Pharmaceutical Statistics, 11:367 – 372, 2012.

H. Southworth and J. E. Heffernan.

texmex: Threshold exceedances and multivariate extremes, 2012.
R package version 1.4.

AstraZeneca