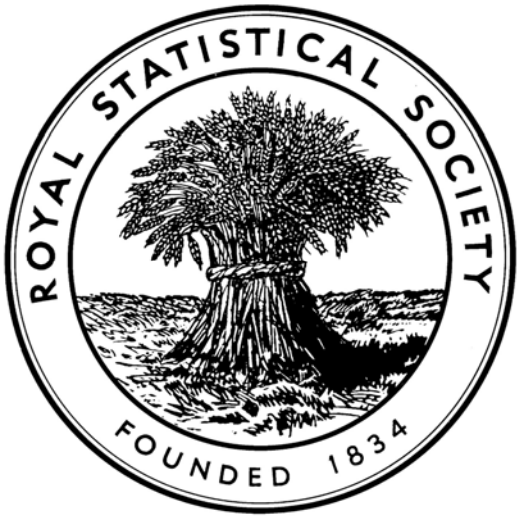
# Report of the Working Party on Statistical Issues in First-in-Man studies




**Royal Statistical Society**
**March 2007**

# Royal Statistical Society
# Working Party on Statistical Issues in First-in-Man Studies

Contents

# PREFACE

## by Tim Holt, President of the Royal Statistical Society

In March 2006 a first-in-man trial took place using healthy volunteers involving the use of monoclonal antibodies. Within hours the subjects had suffered such adverse effects that they were admitted to intensive care at Northwick Park Hospital.

In April 2006 the Secretary of State for Health announced the appointment of Professor (now Sir) Gordon Duff, who chairs the UK's Commission on Human Medicines, to chair a scientific expert group on phase 1 clinical trials. The group reported on 7 December 2006 (Expert Scientific Group on Clinical Trials, 2006a).

Clinical trials have a well established regulatory basis both in the United Kingdom and worldwide. Trials have to be approved by the regulatory authority and are subject to a detailed protocol concerning, among other things, the study design and statistical analyses that will form the basis of the evaluation. In fact, a cornerstone of the regulatory framework is the statistical theory and methods that underpin clinical trials.

As a result, the Royal Statistical Society established an expert group of its own to look in detail at the statistical issues that might be relevant to first-in-man studies. The group mainly comprised senior fellows of the Society who had expert knowledge of the theory and application of statistics in clinical trials. However the group also included an expert immunologist and clinicians to ensure that the interface between statistics and clinical disciplines was not overlooked. In addition, expert representation was sought from Statisticians in the Pharmaceutical Industry (PSI), an organisation with which the RSS has very close links.

The output from the Society's expert group is contained in this report. It makes a number of recommendations directed towards the statistical aspects of clinical trials. As such it complements the report by Professor Duff's group and will, I trust, contribute to a safer framework for first-in-man trials in the future.

# E. Executive Summary

The first in man trial of TGN1412, which began on 13 March 2006, and in which six healthy volunteers suffered immune reactions with severe and in some cases long term consequences, raised serious issues for the future conduct of clinical trials in new treatments. Because, self evidently, the purpose of clinical trials is to produce information and because statistics is the science of information, many of the issues raised are statistical. Accordingly, the Royal Statistical Society considered it only right and proper to constitute a working party to consider statistical issues in first-in-man studies. This is the report of the working party.

## *E.1. Terms of Reference*

The terms of reference of the working party were as follows

"To review statistical design considerations for first-in-man studies with particular reference to monoclonal antibodies and the wider class of new biologicals and biotechnologies.

In particular, to consider the following:

1. What statistical theory, for example as covered by the subjects of decision analysis and experimental design, implies about ethical and logical design of first-in-man studies

2. What data are currently available regarding safety of first-in-man studies and what can be done to facilitate their use in guiding trial design

3. What statistical practice, in particular as regards future appropriate collection and dissemination of data, can contribute to improving ethical decision making in first-in-man studies."

## *E.2. Membership of the working party*

The membership of the Working Party is given in Appendix 1.

## *E.3. Outline of the report*

Chapter 1 is mainly about history and chapter 2 about current practice and related issues. Chapter 1 gives an account of the events leading up to the TGN1412 trial of 13 March and its subsequent reporting. Chapter 2 gives some general discussions regarding first-in-man studies, how they are currently conducted and the working party's attempts to obtain data under the Freedom of Information Act regarding the risks attendant on such studies. These two chapters may be omitted by readers interested in technical recommendations.

Chapter 3, which is a special chapter contributed by Professor Sir Peter Lachmann, FRS, FMedSci, covers immunological matters relevant to the conduct of the trial of TGN1412 and of any future analogous trial.

Chapter 4 covers ethical issues in conducting clinical trials from a statistical perspective, and in particular risk and the conveying of risk from the point of view both of society and of trial subjects.

Chapter 5 considers safety databases: what data are currently available and what steps should be taken to improve the collection and sharing of information on risk of first-in man-studies.

Chapter 6 is concerned with pre-clinical studies and in particular how they should be designed, analysed and used in order to guide the design of first-in-man studies.

Chapter 7 concerns the design and conduct of first-in-man studies. It is the most technical of the chapters and applies statistical theory to various practical matters in designing trials. The relationship between the design, the intended analysis and the purpose of first-in-man studies is stressed.

The list of recommendations made by the working party is given below. Recommendations include generic issues (E.4.1 to E.4.3), preparatory work prior to first-in-man studies (E.4.4 to E.4.9); the design of trials and content of protocols (E.4.10 to E.4.16); agreed norms for risk information sharing for social good (E.4.17 & E.4.18) and reporting standards (E.4.19 to E.4.21). Cross-references are made to particular sections of the report where these are discussed in greater detail. Where such cross-references are not given, the recommendations represent general views that have emerged as a consensus of the working party in reviewing the chapters of the report.

## *E.4. Recommendations by the working party*

**E.4.1.** Governance arrangements for NHS Research Ethics Committees ensure that statistical expertise relevant to research is brought to bear. Since the Central Office for Research Ethics Committees has advised that ethics committees may generally rely on the Medicines and Healthcare products Regulatory Agency (MHRA) to assess the safety of medicinal trials, the MHRA should make at least equivalent provision for statistical expertise.

**E.4.2.** All participants in first-in-man studies in healthy volunteers should be insured. (See 4.3.6.)

**E.4.3.** Studies wherein even the remote possibility of such complications as cytokine storm has been foreseen should be carried out only in hospitals with full facilities for giving tertiary care. (See 1.1.1 & 3.2.)

**E.4.4.** Before proceeding to a first-in-man study, there should be:

- Quantitative justification of the starting dose - based on suitable pre-clinical studies and relevant calculations. (See 6.3.4 & 7.5.1.)
- A priori assessment of the risk level for the recommended study dose(s). (See 7.5.2.)
- Appraisal of the uncertainty about these recommendations (See 1.3, 4.3.1 & 6.3.2.)

**E.4.5.** Sponsors, and MHRA assessors independently, should classify the finally proposed doses of an investigational medicinal product by level of a priori assessed risk (Low, Medium High), having regard to the confidence in pre-clinical data (See section 6.3.3.). High a priori risk would rule out participation by healthy volunteers.

**E.4.6.** Besides first-in-man studies of any monoclonal (regardless of intended target) or of other novel molecules that target the immune system via a novel mechanism, a precautionary approach to study design is appropriate for any experimental medicine which is 'first in class' (and therefore High a priori risk level) or whose a priori assessed dose-specific risk is High.

**E.4.7.** Assessment described in E.4.3 above should be available in a separate document - to be provided to research ethics committee, study participants, and insurers. (See 4.3.2.)

**E.4.8.** Crude inter-species scaling may be inadequate for establishing the initial dose of 'biological' treatments in man. In vitro studies using human cells will be necessary to establish the avidity of the ligands for their targets and assist in dose calculation.(See chapter 3, and sections 6.3.1, 6.3.3 & 6.4.1.)

**E.4.9.** Unless arguments have been provided that the risk is so low that simultaneous treatments is acceptable, in order to allow early evidence of toxicity to halt the trial without risk to subsequent subjects, a proper, or sufficient, inter-administration interval needs to be proposed and observed. (See 1.1.1, 4.3.3 & 7.5.4.).

**E.4.10.** First-in-man study protocols should provide:

- Justification of the proper interval between administration to successive subjects. (See 7.5.4 & E.4.11.)
- Justification of the dose steps the trial will use. (See 7.5.7.)
- Operational definition of 'safety' if investigating safety and tolerability
- Delay between receiving biomarker or other laboratory results which determine 'safety' and having obtained the relevant biological sample. (See 1.1.1 & 7.5.6.)
- Prior estimates of the expected number (or rate) of adverse reactions by dose, especially those serious enough to raise questions about 'safety'). (See 1.1.2 & 4.3.1.)

**E.4.11.** Appropriate sample sizes for first-in-man studies can be better justified statistically - rather than by mere custom and practice – when 'safety' has been given an operational definition. (See 1.1.1.)

**E.4.12.** First-in-man study protocols should discuss their chosen design and its limitations together with the implications for analysis. For example, if an unequal allocation between treatment and placebo per dose step is chosen, this affects the ability of the data-safety monitors to assess tolerability most efficiently before proceeding to a further dose-escalation step. (See 1.1.1 & 7.5.8.)

**E.4.13.** First-in-man study protocols should describe their intended analysis in sufficient detail to allow protocol reviewers (and the independent research ethics committee) to determine if the objectives, design and proposed analyses are compatible. (See 7.5.10 & 7.5.12.)

**E.4.14.** The design of first-in-man trials and the analysis of the data should reflect realistic models of the pharmacokinetic data. (See 7.5.11.)

**E.4.15.** The plan for blood sampling and analysis and observation of vital signs should be based on information from pre-clinical studies. (See 7.5.5.)

**E.4.16.** For first-in-man studies, the standard of informed consent to be observed is 'open protocol, hidden allocation' - that is, all aspects of the trial design shall be shared with subjects to be recruited. (See 4.3.4.)

**E.4.17.** Public debate and research are needed about the maximum acceptable level of risk for first-in-man studies in healthy volunteers and about whether there should be risk-adjusted remuneration of healthy volunteers. (See 1.1.1, 1.1.2 & 4.3.7)

**E.4.18.** Competent drug regulatory authorities should provide a mechanism for the pharmaceutical industry to collect and share data on serious adverse reactions in first-in-man studies – to improve a priori risk assessment. (See 5.8.1 & 5.8.2.)

- For example, separate syntheses of study designs and of the occurrences of predicted, theoretical and unprecedented harms – either as serious adverse events or distributional changes in biomarkers – should be considered for healthy volunteers and for patients, by type and novelty of compound, and by a priori assessed level of risk.
- In particular, for the UK, MHRA should report annually on the designs of, and serious adverse events (whether for the first exposed cohort or at a dose-escalation step) in, first-in-man studies in healthy volunteers (versus patients) that involved administration a biological/biotechnology; and for those that involved a chemical compound.) (See 2.4, 5.8.1 & 5.8.2.)
- The MHRA should also take responsibility for maintaining a central registry of participating volunteers in the UK.

**E.4.19.** Statistical reporting of pre-clinical studies should be improved to be comparable to the requirements by the International Conference on Harmonisation for the reporting of clinical trials: see ICHE9. (See 6.4.6.)

**E.4.20.** Greater use should be made of numerical, as opposed to verbal, descriptions of risk and statistical variation in the submissions made to, and accepted by, competent drug regulatory authorities. (See 4.3.7.)

**E.4.21.** Mock applications to Competent Authorities convey expected standards. They should be revised to: a) be in conformity with the preceding recommendations on the statistical reporting of pre-clinical studies, b) require always that a proper inter-administration interval between successive subjects is both specified and justified, c) specify the waiting time for laboratory-based results which pertain to 'safety': (See E.4.11 above and 2.4.)

## *E.5. Acknowledgements*

The working party gratefully acknowledges the help of Dr Martyn Ward, Head, Clinical Trial Unit, MHRA and Dr Mark Lewis, Director of Clinical Governance and Research, NWL, NHS London, in providing information requested under the Freedom of Information Act.

# 1. Introduction: History and Remit

In this chapter we recount the history of the first-in-man trial of TGN1412 both up to and after the events of 13 March 2006. We also explain the reason for the formation of the Royal Statistical Society working party and its remit.

## *1.1 TGN1412:*

### *1.1.1. Recent event-history*

In November 2005, the Academy of Medical Sciences (AMedSci) published its report on *Safer Medicines* (The Academy of Medical Sciences, 2005), to which several members of UK's Medicines and Healthcare products Regulatory Authority (MHRA) contributed in a personal capacity.

On monoclonal antibodies, AMedSci (The Academy of Medical Sciences, 2005) warned that toxicological approaches applied to chemical drugs were of only limited relevance to biologicals, which are molecularly complex and often heterogeneous in composition, with the potential for greater batch-to-batch variation in composition and for vulnerability to contamination [p27]. Moreover [p16], a drug may be so selective for the human target that it has little effect in the animal species used for safety assessment, "as is the case for most monoclonal antibodies and some organic chemicals."

On 23 December 2005, valid application was made to MHRA by TeGenero AG to conduct a 'First Time in Man' healthy volunteer study of intravenous TGN1412, which constituted a new class of monoclonal antibody with a stimulatory mode of action on target cells of the human immune system. For TGN1412 the target was a subset of T cells - the regulatory Tcells (TRegs). The selectivity of TGN1412 for TRegs was based on experimental studies in rodents and does not follow obviously from its specificity, since its ligand, CD28, is present on the majority of T-cells. The failure to explore adequately the basis of this surprising selectivity may be the root cause of the problems with this particular trial; and would not apply to other monoclonal antibody trials where the anticipated effects clearly follow the ligand specificity.

On or after 24 January 2006, the TGN1412 'First Time in Man' Phase 1 protocol in healthy volunteers was approved by MHRA, whose assessors' reports were dated 6, 23 and 24 January 2006.

The medical assessor queried neither TGN1412's administration to healthy volunteers (versus patients), nor the time-intervals between a) the initial and subsequent intravenous infusion of TGN1412 to other volunteers in the same cohort, and b) infusion and availability of immunological monitoring data. Nor did the medical assessor query the 'appropriate measures' to deal with the possibility of cytokine release storm. Benefit to risk ratio for healthy volunteers was described as 'favourable'.

On 4 January 2006, valid application was made for ethical approval of the TGN1412 protocol to Brent Medical Ethics Committee (BMEC). The application was reviewed in committee on 23 January 2006 when, subject to a complete response (received on 9 February 2006) to its request (of 31 January 2006) for further information, BMEC would be content to give a favourable ethical opinion. It delegated authority to the Alternate Vice-Chair accordingly.

The further information requested by BMEC included explanation of: whom the researchers would consult if they detected T-cell abnormalities (unspecified whether at screening or post-

randomisation); how the seemingly high payment to volunteers had been calculated (to allay concerns about induced recruitment); whether the researchers would screen for carriage of tuberculosis and for HIV seroconversion (because Parexel's volunteers included travellers from South Africa). BMEC also requested a report from, and short curriculum vitae (CV) of, a *clinical* immunologist who should review the toxicology report on TGN1412. The information Sheet was too technical and needed to be simplified (see later: 2.6)

The application to BMEC had stated, and volunteers were also to be told, that: "Expert advice from immunologists has been sought in designing the protocol to minimise your risks, including a robust screening process that takes into account your immune status, and repeated thorough assessment of immune function". A Parexel statistician had also been involved in writing and review of the TGN1412 protocol. The number of participants was explained thus (at A51 in application): 'The total sample size of 32 subjects is not based on a formal statistical assessment. However, this number of subjects is considered sufficient to achieve the objectives of the study. Each group will consist of eight subjects of whom six will receive active drug and two will receive placebo'.

Of three immunologists' responses for BMEC (identities and CVs withheld), only the response on TeGenero-headed notepaper directly addressed design aspects:

> "Therefore, the risk of inducing a clinically apparent cytokine release syndrome (CRS) by administration of TGN1412 in humans is considered to be low, although it cannot be completely excluded. In order to ensure maximum safety of treated individuals, in particular during the early phase of the TGN1412-HV trial, subjects will be monitored closely for pro-inflammatory cytokine levels and clinical signs of CRS.
>
> In order to detect potential autoimmune effects induced by TGN1412 in humans, absolute and relative lymphocyte counts, B- and T- lymphocyte differentiation and activation markers as well as immunoglobulin levels, rheumatoid factor and ANA antibodies will be monitored during the proposed phase 1 clinical trial . . . ".

***CRS was clearly, and explicitly, anticipated; but*** "the risk . . . is considered to be low". Nonetheless, BMEC did not query that eight healthy volunteers would be dosed within 2 hours, what the 'appropriate measures' were for managing CRS nor, indeed, a failure to mention adrenaline in managing anaphylaxis.

The information sheet's description of Day 1 explained that, after numerous blood samples to assess your immunological profile, then: "at the allocated time (approximated between 08.00 and 10.00), you will be given your study medication in the form of an intravenous infusion (where the study drug is administered directly into a vein). For practical reasons, the time of dosing will be staggered. You will . . . "

On 14 February 2006, BMEC confirmed its favourable ethical opinion for the by-now MHRA-approved TGN1412 study in healthy volunteers to be conducted by Parexel Clinical Pharmacology Unit (PCPU), a contract research organisation (CRO).

BMEC had specific recognition in appraisal of healthy volunteer studies. It did not have a statistician-member, but membership included other professionals who design research studies and conduct statistical analyses.

On 22 February 2006, PCPU began screening volunteers for participation in the TGN1412 study.

MHRA, BMEC and Parexel knew, but as far as we are aware, study participants were not explicitly informed, that only two (25%) out of eight volunteers (one of participants 001-004; and one of participants 005-008) were to be randomised to placebo.

In March 2006, the legal powers of MHRA and independent research ethics committees in the matter of suspending authorisations for a clinical trial of an investigational medicinal product (IMP) were clarified: please see section 1.2.

On 13 March 2006, 'First Time in Man' intravenous administration of TGN1412 began at 08.00 hours with the sixth intravenous administration at 09.00 hours on 13 March 2006 (Expert Scientific Group on Clinical Trials, 2006b). Clinical symptoms began within one to two hours of TGN1412's administration (Suntharalingam et al., 2006). By midnight on 13 March 2006, all six males who had received TGN1412 intravenously had been admitted to Northwick Park Hospital in need of intensive care.

On 14 March 2006, there was communication between BMEC and Parexel/TeGenero about TGN1412 volunteers' admission to intensive care.

Also, on 14 March 2006, MHRA wrote to Parexel suspending its authorisation for the TGN1412 study. Copy of MHRA's letter was received by BMEC on 15 March 2006.

#### *1.1.2. Media reporting*

On 14 March 2006, print and UK broadcast media first broke the news that six participants in a first-in-man study involving healthy volunteers had been admitted to intensive care in a district general hospital, Northwick Park, within hours of intravenous administration of a new experimental medicine.

On 15 March 2006, interviewees on BBC Radio 4 *Today* programme stressed a reassuring rarity of serious adverse events in first-in-man studies, and the need for volunteering to continue. Neither journalists nor interviewees commented that six, more or less simultaneous and identical, serious adverse events in a first-in-man study were unprecedented. Unsurprisingly, and properly, the police had been informed lest the study had been sabotaged. No-one initially queried (Hawkes, 2006b; Lister, 2006a) whether – even in the event of sabotage – a more risk-precautionary study design could have prevented exposure to harm of so many healthy volunteers.

Within 48 hours, the *Times* and other print journalists were reporting that the experimental medicine, TGN1412, was a monoclonal antibody – which raised further questions about the propriety of having enlisted healthy volunteers (rather than patients) (The Academy of Medical Sciences, 2005).

On 16 March 2006, the *Times* (Hawkes, 2006b) cited the Textbook of Pharmaceutical Medicine (Posner, 2005) on size of cohorts and how to minimise risk (including by ensuring sufficiently long inter-administration intervals). Initially, calls for publication of TGN1412 protocol, and its release to volunteers' families, were resisted. This injunction was soon rightly reversed in the wider public interest.

By the weekend of 18 March 2006, it became clear that TGN1412, due to its stimulatory mode of action, was unusual even among monoclonal antibodies. Well-informed immunologist and

triallist commentators were prudently calling for a moratorium on first-in-man studies of monoclonal antibodies until their experimental designs had been re-assessed in the shadow of TGN1412 but also reminded readers about the potential benefits from new drug discoveries (Calne, 2006; Editorial, 2006b; Hawkes, 2006b; Irving, 2006; Jones & Gordon, 2006; Lee, 2006; Lister, 2006b; McKie & Revill, 2006; Revill, McKie, & Hill, 2006; Rogers & Oakshot, 2006; Rogers, Woods, & Deer, 2006).

On 6 April 2006, the *Times* (Lister, 2006b) reported that, to assuage concerns, MHRA had reviewed its protocol-approvals since May 2004 for 40 first-in-man Phase 1 studies of monoclonal antibodies. Thus, contrary to initial and well-publicised reassurances, which were based mainly on the safety record of first-in-man studies of chemical compounds, the empirical basis for corresponding claims in respect of new biologicals or biotechnologies – especially of monoclonal antibodies in healthy volunteers – was severely limited in March 2006.

A number of editorials and personal views in professional journals (Goodyear, 2006; Mandeville, 2006; Mayor, 2006; Schneider, Kalinke, & Lower, 2006; Wood & Darbyshire, 2006) followed, to some of which we refer and which anticipated many recommendations to be made formally at a later date (Expert Scientific Group on Clinical Trials, 2006b).

*1.1.3. Implications for MHRA and for independent research ethics committees*

On 5 April 2006, MHRA reported its interim findings (subsequently confirmed) that TGN1412, as administered, had been properly produced; and that its administration complied with the MHRA-approved protocol.

On 5 April 2006 also, MHRA released the TGN1412 protocol, Investigator Brochure (IB), Investigational Medicinal Product Dossier (IMPD), and Assessment Report. Some of the biological rationale, planned measurements to have been made on volunteers' samples, references, laboratories' and other names – including of assessors - were blanked out by reason of commercial or other confidentiality. See (MHRA Press office, 2006). Freedom of Information request(s) had been received such as NB122/6 from the BBC on 17 March 2006.

On 20 April 2006, Ministers announced the appointment of Professor (now Sir) Gordon Duff, who chairs the UK's Commission on Human Medicines (CHM), to chair an Expert Scientific Group (ESG) on Phase 1 Clinical Trials.

As of 1 May 2006, MHRA adopted a precautionary approach *to first-in-man (Phase 1) trials of any monoclonal antibody (regardless of intended target) or other novel molecules that target the immune system via a novel mechanism*: please see section 1.2.

On 23 May 2006, ESG held the first of six meetings prior to submitting an interim report to Ministers on 20 July 2006.

In June 2006, with prior authority from Council, a Royal Statistical Society Working Party, chaired by Professor Stephen Senn, was convened to review statistical aspects of first-in-man studies: please see section 1.4.

On 25 July 2006, ESG's interim report and supporting papers were released by Ministers for public consultation until 14 September 2006. The Royal Statistical Society's (RSS) Working Party contributed its draft recommendations, and was represented at a subsequent meeting of ESG on 24 October 2006 to discuss them.

On 3 August 2006, RSS Working Party lodged a Freedom of Information request with MHRA about its approved-designs (and associated serious adverse event reporting) for Phase 1 studies in UK, see section 2.3.

On 14 August 2006, the report by Suntharalingam et al. on cytokine storm (Suntharalingam, et al., 2006) in a Phase 1 trial of anti-CD28 monoclonal antibody TGN1412 was published online.

In late September 2006, both BBC Radio 4 and Channel 4's Dispatches produced programmes about ESG's report and TGN1412 volunteers; and, on 14 October 2006, an interim viewpoint by Dutch regulators was published in (Kenter & Cohen, 2006).

On 7 December 2006, ESG's final report and supporting papers were published (Expert Scientific Group on Clinical Trials, 2006a).

#### *1.1.4. Global concern*

Failure of an MHRA-approved and ethically-approved protocol in the TGN1412 study to minimise harm to (healthy) volunteers, and particularly to volunteers C, D, E, F, G in **Table 1.1**, made news around the world.

Internationally, it should have prompted sponsors and clinical investigators in receipt of relevant 'first-in-man' approvals for Phase 1 studies, particularly of biologicals or biotechnologies, to refer them back to both the Competent Authority, which is MHRA in UK, and to relevant independent research ethics committee for re-consideration in the aftermath of TGN1412.

Competent authorities and independent research ethics committees should proactively have written to sponsors to institute such reviews, and – according to their legal powers (see section 1.2) – put a moratorium on previous permissions until the reviews were speedily conducted. In UK, this action was instituted (Lister, 2006b) by its Competent Authority (MHRA) in respect of higher risk IMPs (see 1.2.1) because only MHRA has legal powers to suspend clinical trial authorisations (CTAs) for IMPs. A Freedom of Information request to the MHRA elicited that six first-in-man studies were identified which had been approved or were pending and had not yet started. In all six, the sponsor agreed voluntarily to suspend the relevant CTA or application until CHM had provided expert advice.

In the aftermath of TGN1412, the Medical Research Council, for example, immediately reviewed study designs that it had funded in a research call in 2005 on Experimental Medicines, so that relevant grant-holders could consider the implications that any design failures in the TGN1412 protocol might have.

MHRA had instigated speedy reviews of the quality of the administered batch of TGN1412 and of protocol-adherence in its administration at PCPU. The latter properly accorded with TGN1412's MHRA-approved protocol. Neither review made untoward findings.

Regulatory or scientific short-comings on the other hand, as Goodyear (Goodyear, 2006) had noted, needed independent appraisal, which Professor Duff's appointment to chair an Expert Scientific Group (ESG) on Phase 1 Clinical Trials signified.

Despite immense public interest, TGN1412-administered volunteers' individual sequences of baseline data, administration and adverse event onset times, harm reduction measures in PCPU,

and time-specific key immunological data were not fully disclosed by ESG in order not to pre-empt imminent peer-reviewed publication in the New England Journal of Medicine (Suntharalingam, et al., 2006). See also (Expert Scientific Group on Clinical Trials, 2006a, 2006b)

**Table 1.1: Event times for 8 healthy volunteers in TGN1412 study 13 March 2006**

| Healthy volunteer (see ESG: baseline data) | Randomised to TG = TGN1412 | Time of intravenous administration | Time after dosing to critical care transfer |
|---|---|---|---|
| A | TG 8.4mg | 08:00 | 16 hrs |
| B | Placebo | 08:10 | |
| C | TG 6.8mg | 08:20 | 15hrs 30mins |
| D | TG 8.8mg | 08:30 | 16hrs |
| E | TG 8.2mg | 08:40 | 12hrs |
| F | TG 7.2mg | 08.50 | 16hrs |
| G | TG 8.2mg | 09:00 | 16hrs |
| H | Placebo | 09:10 | |

From ESG's description, cytokine results may not have been available in PCPU on 13 March 2006.

## *1.2 Important UK changes*

Changes in section 1.2.1 are interim pending ESG's final recommendations and decisions thereon by Ministers.

### *1.2.1 MHRA changes as of 1 May 2006 in the aftermath of TGN1412*

MHRA adopted a precautionary approach to: *first-in-man (Phase 1) trials of any monoclonal antibody (regardless of intended target) or other novel molecules that target the immune system via a novel mechanism*.

Such trials will not be sanctioned by UK's Competent Authority without having had additional expert opinion from Commission on Human Medicines (CHM) on whether cytokine storm, as in TGN1412, may be repeated in relation to those substances.

Accordingly, *two questions have been added* to Phase 1 study applications to identify:
i) first-in-man studies, and
ii) if their IMP is one to which MHRA's precautionary approach applies.

Moreover, sponsors are encouraged to seek CHM's advice *prior* to making applications for either clinical trial authorisation (CTA) or independent research ethics committee (REC) approval so that CHM's opinion can influence:
a) final study design,
b) how potential risks are described in the patient/volunteer information sheet (together with implications for insurance and indemnity provision), and
c) measures in place to minimise potential risks (for example, stopping rules, emergency procedures, intensive care facilities).

In addition,
d) CTA authorisation *should precede submission* to independent REC so that REC can have access to CHM's expert opinion and understand, accordingly, how the study protocol was modified to minimise risks to participants.
e) *Precautionary approach is retrospective* for relevant first-in-man studies with CTA and REC authorizations, but which had not started prior to May 2006.

### *1.2.2 Independent research ethics committee changes as of March 2006*

Independent RECs have been:
f) advised in relation to MHRA assessments, but
g) alerted to their limited legal powers in respect of CTAs for IMPs in Phase 1 trials, and
h) confirmed in their powers in respect of other research studies.

In particular,
f). All National Health Service (NHS) RECs were advised that: *'The ethics committee may generally rely on the MHRA to assess the safety of medicinal trials. It is not required to undertake its own safety assessment or seek expert advice on safety issues from scientific referees'*.

g). Based on new legal advice about Medicines for Human Use (Clinical Trials) Regulations 2004, in the case of: *a clinical trial of an investigational medicinal product (CTIMP), the REC may not suspend or terminate its opinion.*

Power to take action in the event of emerging concerns lies entirely with MHRA by suspending or terminating CTA under Regulation 31. The REC can only action any concerns it has by writing to, or emailing, the Head of the Clinical Trials Unit at MHRA. But, the REC continues to have a statutory duty to give an opinion on substantial amendments notified by the sponsor, and may give a favourable or unfavourable opinion.

h). Confusingly, this policy change in respect of clinical trials of IMPs *does not affect* the REC's ability to suspend or terminate its opinion on any other research study in accordance with Governance Arrangements for NHS Research Ethics Committees and standard operating procedures.

### *1.2.3. Has statistical assessment changed as a consequence?*

A concern is that statistical assessment is not an obligation on MHRA in the granting of CTAs, whereas Governance Arrangements for NHS Research Ethics Committees sought to ensure it, for example:

- (6.4) The "expert" members of the committee shall be chosen to ensure that the REC has the following expertise . . . the third category of which is 'statistics relevant to research'.

- (9.10) If the committee is of the opinion that the prior scientific review commensurate with the scale of the research is not adequate (including adequate statistical analysis), it should require the applicant to re-submit the application having obtained further expert review.

- (9.13) Requirements for a favourable opinion . . . a) the appropriateness of the study design in relation to the objectives of the study, the statistical methodology (including sample size calculation where appropriate), and the potential for reaching sound

conclusions with the smallest number of research participants; b) the justification of predictable risks and inconveniences weighed against the anticipated benefits for the research participants, other present and future patients, and the concerned communities; . . . h) the manner in which the results of the research will be reported and published.

### *1.3. Interim report of Expert Scientific Group.*

Background papers to ESG's interim report, including by the Association of the British Pharmaceutical Industry and Faculty of Pharmaceutical Medicine of the Royal Colleges of Physicians, emphasised that when the science relating to the mechanism(s) of action of a novel experimental medicine in terms of both desired and potential untoward effects is poorly understood by reason of complexity or novelty, it is incumbent on reviewers – be they sponsor, regulator, or REC - to seek expert advice and to review data from novel compounds with similar or related mechanism of action.

Cautious administration is not a new requirement (Hawkes, 2006a; Posner, 2005), is certainly appropriate for a very novel new experimental medicine which is 'first in class', but has wider application dependent on the a priori assessed risk which may itself be dose-related. As there is no direct benefit for healthy volunteers (other than remuneration, learning about research, and altruism), the risks to them must be minimal. For example, there should be a high level of confidence in the predicted safety of the IMP based on pre-clinical data.

As AMedSci had pointed out, these pre-conditions seldom apply for monoclonal antibodies. Also, the Medical Research Council's submission to ESG recognised that the development of novel therapeutic agents - particularly those with new mechanisms of action, high species-specificity of action, or directed at immune system targets - inherently possess a high degree of *uncertainty* about risk; and that this uncertainty has to be managed by appropriate independent scientific consideration of preclinical data, appropriate design of the initial clinical studies, and conveyed in the information supplied to participants, clinical investigators, trialists, and others who may be involved.

ESG's interim and final reports (Expert Scientific Group on Clinical Trials, 2006a, 2006b) made 22 recommendations, and touched on such statistical issues as:

- synthesis of pre-clinical evidence with a view to a priori risk classification;
- whether it is appropriate to study healthy volunteers;
- choice of starting dose and its escalation;
- study design with particular reference to initial versus dose-escalation cohort sizes, and allocation ratios;
- information for healthy volunteers (including on risk classification, cohort sizes and allocation ratios);
- uses of pre-clinical databases on biologicals or biotechnologies;
- database-design, analysis, and access, together with the reporting of Phase 1 study designs, their outcomes and adverse events; and
- training in the skills, including statistical, needed for protocol assessment and risk appraisal.

Additional statistical issues are:

- need for improved statistical reporting standards in submissions made to drug regulatory authorities;
- formal syntheses by drug regulators of safety data in pre-clinical studies submitted to them by *different* sponsors; and
- public debate about risk-adjusted remuneration of healthy volunteers.

### *1.4 Royal Statistical Society's remit*

The Royal Statistical Society (RSS) convened a Working Party:

"To review statistical design considerations for first-in-man studies with particular reference to monoclonal antibodies and the wider class of new biologicals and biotechnologies. In particular, to consider:

- What statistical theory, for example decision analysis and experimental design, implies about ethical and logical design of first-in-man studies

- What data are currently available regarding safety of first-in-man studies, and what can be done to facilitate their use in guiding trial design

- What can statistical practice contribute to improving ethical decision making in first-in-man studies, as regards appropriate collection and dissemination of data."

Clear subject-matter understanding is essential for statistical principles to have practical application, and explains why cross-disciplinary composition was seen as essential for this RSS Working Party. Section 3 deals with relevant, and difficult, immunological concepts that assessment of a first-in-man study design for TGN1412 needed to grapple with.

*Statistics and Statisticians in Drug Regulation*, the report of a previous RSS Working Party (Pocock et al., 1991), was instrumental in the early 1990s in promoting the employment of professional statisticians by UK and other European drug regulatory authorities so that there could be a dialogue of equals with statisticians in the pharmaceutical industry, where statistical science was already well-represented. Regulatory recognition in the 1990s of the key role that statistical science has in the design, analysis and reporting of pharmaceutical research programmes was welcome, as was statisticians' greater representation on drug licensing and appeals panels.

Statisticians' role in the design, analysis and reporting of pharmaceutical research spans the whole range of clinical studies from Phase 1, including first-in-man studies, to major pre-licensing Phase 3 trials for regulatory approval, as well as pre-clinical animal studies and post-licensing pharmaco-epidemiology and pharmaco-vigilance. Drug safety issues can emerge pre-clinically, or only post-licensing (Kramer et al., 2006) when their emergence may call into question past regulatory decisions (Dhaun, Maxwell, & Webb, 2006; Editorial, 2006a; Gunnell & Ashby, 2004; Healy, Herxheimer, & Menkes, 2006; O'Neill, 2002; Whittington et al., 2004). Statisticians' input concerns matters of ethics, science, public health, reporting standards, practicality and common-sense, not just the purely methodological. Statisticians in the pharmaceutical industry, and in regulatory authorities (Food and Drug Administration, 2006; Ioannidis et al., 2004; Kramer, et al., 2006; O'Neill, 2002; The Cochrane Collaboration, 2006), thus have a key role in enhancing the scientific and ethical quality of all these phases of pharmaceutical research.

Accordingly, the RSS Working Party's recommendations range from generic issues (E.4.1 to E.4.3), preparatory work prior to first-in-man studies (E.4.4 to E.4.9); the design of trials and content of protocols (E.4.10 to E.4.16); agreed norms for risk information sharing for social good (E.4.17 & E.4.18) and reporting standards (E.4.19 to E.4.21.).

# 2. Some general considerations

In this section we discuss various issues regarding current conduct of first-in-man studies, their role in drug development, what is known about their risks and how these risks could be better communicated. Answers to Freedom of Information requests are summarised. Many of the issues raised here are treated in more detail in subsequent chapters. The knowledgeable reader who is more interested in technical details may wish to skip this chapter.

## *2.1. Interface between pre-clinical and Phase 1 clinical studies*

Experiments should use the smallest number of animals that can clearly answer the question posed, and take every practical step to avoid distress or suffering. Yet, animal studies may lack formal calculation of statistical power or justification of allocation of animals between doses. Festing et al. advocate better experimental design as a means to reducing the use of animals in research (Festing et al., 2002).

Pre-clinical evaluation of safety is by skilled scientists who know about both animal and human pathology. Traditionally, there is heavy reliance on macroscopic and microscopic examination of organs and tissues from at least two species of animals at a range of drug doses and durations.

Drug concentrations and kinetics in blood and tissues are used to calculate the maximum exposure (such as dose per kilogram) in animals that causes no harm, the No Observable Adverse Effect Level (NOAEL). Alternatively, the initial dose in man should be set below the Minimal Anticipated Biological Effect Level (MABEL), or micro-dosing can be considered. See ESG's final report (Expert Scientific Group on Clinical Trials, 2006a).

Administered dose multiplied by the species-specific avidity constant equals effective dose. Avidity constants can usually be worked out in vitro (Kenter & Cohen, 2006): see also chapter 6. An Effective Human Dose (EHD) is usually worked out from this As a rule, human exposure is kept below the EHD, and to no more than one tenth of it if the initial toxicity in animals is of a serious nature (Food and Drug Administration, 2002; Modi, 2006). Safety pharmacology aims to ensure no important functional effects upon vital systems (heart, blood pressure, breathing, consciousness, etc), and – at later phases – aims to provide a mechanistic understanding of adverse events in humans that were *not predicted* from earlier testing.

Pre-clinical evaluation of safety faces three problems: drawing inferences from very high exposures; understanding, or anticipating, the mechanism of toxicity when background knowledge about the biological role of the target receptor or enzyme is limited (see chapter 3); and interspecies comparisons especially 'when a drug is so selective for the human target that it has little effect in the animal species used for safety assessment, as is the case for most monoclonal antibodies and some organic chemicals' (The Academy of Medical Sciences, 2005). *Safer Medicines* saw as vast the potential applications of genomics to safety, and particularly important was the possibility to reduce, or largely avoid, toxicities that have an immunological component as - clearly - animal safety tests are of little value in predicting them.

Particularly critical, and robustly questioned within major pharmaceutical companies, is the decision to proceed from pre-clinical to clinical studies in man. Whether clinical studies can justifiably recruit healthy volunteers to whom there is risk but no benefit – other than reasonable remuneration, learning about research, and altruism – is a key decision.

Transfer of expert knowledge from pre-clinical scientists to the (usually different) team responsible for designing subsequent clinical phases of a pharmaceutical research programme is very important. Pre-clinical scientists' expert knowledge is essential prior information in determining a Phase 1 study design. See chapter 6.

Starting dose is usually a fraction of NOAEL, less than MABEL, or a micro-dose. A proper inter-administration interval serves to minimise risk to the next participant. Pre-set criteria determine when dose escalation should cease (for example: pre-defined upper dose limit; or that some pre-specified number with dose-limiting symptoms of those randomised to a given dose shall rule out administration of next higher dose; or by use of surrogate markers as a safety alert, be they biochemical, haematological or other biomarker). In the light of subject-matter knowledge and prior information from pre-clinical studies, a range of statistical considerations applies in choosing experimental design and stopping rules – and especially so for 'first-in-man' Phase 1 studies (see chapter 7). Risk attends each dose escalation, however.

Phase 1 studies aim to provide reliable dose-specific data on plasma concentrations by duration of exposure. A second Phase 1 study usually considers the effect of repeated dosing (over a specified number of days, say) of volunteers who have been randomised between placebo/active dose level or to dose-escalation pattern; and similar safety measurements are made. Sponsors may run a series of Phase 1 studies, one of which is first-in-man.

### *2.2. Contract Research Organisations*

Contract Research Organisations (CROs), such as Parexel, can be both of value and of concern. Of value because CROs constitute an additional tier of expertise and scrutiny of both protocol and ethics for Phase 1 (and subsequent) studies - with their own Outline Phase 1 Study Protocol for quality-assurance, but of concern if they are not well informed about the investigational medicinal product with the result that confusion or disagreement about responsibility ensues.

CROs guard against over-volunteering by having participants' consent that their national insurance and passport numbers be logged on The Over Volunteering Prevention System (TOPS, 2006) together with dates of start and end of their study participation. They also seek permission to contact the volunteer's general practitioner**.** An additional check on volunteers' entitlement to NHS treatment may be needed (Expert Scientific Group on Clinical Trials, 2006a). Frequently asked questions about medical trials and volunteering are advertised at a website (EnterTrials, 2006).

Evidence to ESG (Expert Scientific Group on Clinical Trials, 2006a, 2006b) from the Association for Human Pharmacology in the Pharmaceutical Industry was that CROs currently run up to 90% of Phase 1 clinical studies in the UK. Of these, one third have sponsors in the UK, one third have sponsors from elsewhere within Europe, and one third from outside of Europe, mainly Japan. Annual income (howsoever defined) in the UK from Phase 1 studies was reckoned at around £170 millions.

Based on the recent rate of MHRA approvals of Phase 1 studies (see below), and assuming that 90% of approved studies go ahead, CROs may be responsible for running 500 Phase 1 clinical studies each year in the UK, 10% of which (see below) may involve a biological/biotechnology.

*If* mean income per CRO-conducted Phase 1 clinical study were around £340,000 (that is: £170 millions/500), then even individual reimbursement of £1,000 to £1,500 for an average of 40 healthy volunteers per study represents less than *20%* of mean income per CRO-conducted

Phase 1 study. BMEC, which had approved 52 Phase 1 studies in healthy volunteers from 1 April 2004 to 31 March 2006 (answer to Freedom of Information request), had commented that reimbursement of £2,000 per volunteer for full compliance with the requirements for TGN1412's protocol was unusually high by their experience.

### *2.3 Freedom of Information request to MHRA on designs of, and serious adverse event reporting in, Phase 1 studies in healthy volunteers*

Background: Properly reported information on the design of Phase 1 clinical studies in healthy volunteers and on the safety of participants according to type of new experimental medicine was lacking from ESG's interim report. A voluntary organisation for the self-regulation of commercial or academic Phase 1 units was set up after the deaths of two healthy volunteers in studies in Dublin and Cardiff. It provided headline information (only) to ESG on serious adverse events in healthy volunteer studies from 1 January 1992 to 31 December 2000 (nine years) (Expert Scientific Group on Clinical Trials, 2006a, 2006b): 171 serious adverse events had apparently occurred in 81,471 healthy volunteers. (See, however, section 5.1 for a discussion of data.)

However, the number (and percentage) of studies which had reported serious adverse events in healthy volunteers randomised to the IMP (versus placebo) was not reported. Nor was ESG given information on whether the serious adverse events had occurred in the initially-exposed cohort, or upon dose escalation. ESG had no information either about the number of volunteers randomised to IMP (versus placebo) in the affected cohort. These are basic questions for analysts to address. Moreover, answers are needed separately by type of IMP: biological/biotechnology versus chemical compound.

*If* the number of healthy volunteers per Phase 1 study was usually around 10 or 20 or 40, and *if* most studies in healthy volunteers would stop at the first report of serious adverse event(s), *then* the above information for the 1990s could be consistent with 150 Phase 1 studies in healthy volunteers having reported one or more serious adverse events in a total of 2,000 (*if* 40 healthy volunteers on average) to 8,000 Phase 1 studies. By implication, the 1990s' per-Phase 1 study risk of serious adverse event could have been as high as 1.9% to 7.6% (150/2,000) for healthy volunteers.

*MHRA era:* From 1 May 2004, MHRA assumed responsibility for the central regulation of Phase 1 clinical studies in the UK. ESG's background papers (Expert Scientific Group on Clinical Trials, 2006a, 2006b) gave the tally of Phase 1 trials authorised by MHRA "since September 2004" (to, we assume, 31 May 2006: 20 months) as in Table 2.1.

**Table 2.1. Phase 1 trials authorised by MHRA September 2004 to end May 2006**

| Type of Investigational Medical Product (IMP) | MHRA-approved protocol in | | |
|---|---|---|---|
| | Patients (% for type) | Healthy volunteers (% for type) | Total: per month |
| Biological/biotechnology | 26 (28%) | 66 (72%) | 92: 4.6 per month |
| Chemical | 82 ( 9%) | 842 (91%) | 924 : 46.2 per month |

Notice that MHRA-approved protocols for Phase 1 clinical studies of a chemical IMP outnumbered those for a biological/biotechnology by 10:1. Moreover, a higher proportion at 28% of Phase 1 studies of a biological/biotechnology was designed to recruit patients rather than healthy volunteers (versus 9% for chemical IMPs).

*If* we may reasonably assume that 10% of MHRA-approved protocols in healthy volunteers are not realised (by reason of reluctance on the part of clinical investigator, REC, or sponsor), then only three per month in UK actually involve a biological/biotechnology. *At least one* such protocol (TGN1412) in healthy volunteers has been associated with serious adverse event(s) (six, in fact) out of *at most 60* implemented MHRA-approved protocols for non-chemical IMPs, a per-Phase 1 study risk in healthy volunteers of at least 1/60 (1.7%).

**Table 2.2** illustrates that knowledge of the MHRA-approved Phase 1 study designs matters for evaluating risk to healthy volunteers recruited into the initial cohort for a Phase 1 clinical study of biological/biotechnological IMP. In the three scenarios shown, the risk for those randomised to IMP ranges from 2% to as high as 9%. Prior to formal review, opinion at MHRA was that the design of the TGN1412 first-in-man study was rather typical.

**Table 2.2 Phase 1 study design matters for evaluating risk to healthy volunteers**

| Randomisation ratio in initial cohort = IMP: placebo | Number of healthy volunteers in initial cohort | **Risk Estimator 1** % healthy volunteers in initial cohort with serious adverse event(s) | **Risk Estimator 2** % healthy volunteers in initial cohort **& randomised to IMP** with serious adverse event(s) |
|---|---|---|---|
| 1:1, except for TGN1412 protocol's 3:1 | 2, except for TGN1412's 8 | 6/126 (4.8%) | 6/ 65 (9.2%) |
| 1:1, except for TGN1412 protocol's 3:1 | 8 | 6/480 (1.3%) | 6/242 (2.5%) |
| 1:1 for 30 studies, 2:1 for 29 studies, 3:1 for TGN1412 | 2 for 30 studies, 3 for 29 studies, 8 for TGN1412 | 6/(60 + 87 +8) = 6/155 (3.9%) | 6/(30 + 58 + 6) = 6/ 94 (6.4%) |
| 3:1 | 8 | 6/480 (1.3%) | 6/360 (1.7%) |

(Design assumptions for 60 MHRA-approved Phase 1 studies of biological/biotechnology IMP in healthy volunteers, one of which was TGN1412. Serious adverse events assumed ***not*** to have occurred, except in TGN1412. Not all Phase 1 studies are 'first-in-man', which TGN1412 was, so that 'first-in-man' subset may be lower than benchmark of 60.)

*Freedom of Information request to MHRA:* None of the above plausible design approximations for recent Phase 1 studies which involved a biological or biotechnology represents a 'favourable' benefit to risk ratio in terms of serious adverse events for healthy volunteers (per-Phase 1 study range: from 1% to 5%, and from 2% to 9% if healthy volunteer is randomised to IMP in initial cohort).

However, public and professionals' interest in these risk estimates is such that we should not have to rely on speculation. Data should be marshalled by MHRA. By the time of its interim report (Expert Scientific Group on Clinical Trials, 2006b), however, ESG had not had access to the MHRA-approved First in Man study designs, nor to serious adverse event numerators and denominators pertaining to the initial cohorts, let alone for**:**

a) all Phase 1 studies of healthy volunteers that involved administration of a biological/biotechnology, and
b) 10% random sample of Phase 1 studies of healthy volunteers that involved administration of a chemical compound.

Corresponding data on serious adverse events are required also in respect of later, dose-escalated cohorts in the same studies. Design information (and serious adverse event reporting) is particularly important for First-in- Man studies.

Therefore, on 3 August 2006, the Royal Statistical Society's Working Party lodged a Freedom of Information request with MHRA that the actual data be marshalled. Our request was initially denied on 18 August 2006 on grounds of cost, but consequent to an appeal, there was agreement on 24 October 2006 that the Head of MHRA's Clinical Trials Unit would abstract the requested information for MHRA-approved first-in-man studies of biological/biotechnology IMPs and for a 10% sample of chemical IMPs.

The fundamental reason that the requested data were costly for MHRA to marshal is because their databases on Phase 1 study designs and on serious adverse events, which should be capable of being aligned programmatically, cannot be because they lack common identifiers. In particular, prior to June 2006, there was no flag for 'first-in-man' Phase 1 studies. Secondly, there is no link between the clinical trials and safety databases to provide integrated reports of the sort suggested. The data requested under Freedom of Information should allow experimental designs to be compared between uneventful 'first-in-man' Phase 1 studies in healthy volunteers and those reporting serious adverse events, see Table 2.3. The request also helped to identify the links, and analysis protocols, that are a future requirement the MHRA.

In response to RSS's Freedom of Information appeal, Dr Martyn Ward, on behalf of MHRA, extracted summary information on the designs of first-in-man studies approved by MHRA during the first three months of 2005 for chemical compounds (55 studies: 23 in patients and 32 in healthy volunteers) and in the 30 months from May 2004 to October 2006 for biologicals (32 studies: 15 in patients and 17 in healthy volunteers). Dr Ward also provided information about suspected serious adverse reactions (SUSARs) which MHRA was aware of as having occurred in these approved studies. (Please see www.rss.org.uk for Freedom of Information Table on 87 MHRA-authorised first-in-man study designs.)

SUSARs were reported for four of the 55 first-in-man studies of chemical compounds. All four were non-placebo-studies in patients with cancer, and designed to test IMP until toxicity: a single SUSAR was reported for three studies (in first cohort (1) or intermediate cohorts (2)), while the fourth reported four SUSARs in the 3$^{rd}$, 5$^{th}$ and 8$^{th}$ cohorts of a study which was originally designed to recruit up to seven cohorts only.

SUSARs were reported for two of the 32 first-in-man studies of biologicals: the placebo-controlled TGN1412 study in healthy volunteers, which reported SUSARs in each of its six volunteers who received IMP in the first cohort of eight subjects; and a placebo-controlled study in patients with cystic fibrosis, which reported a single SUSAR in the multiple dosing phase of its first cohort of six patients, four of whom received IMP.

For the 34/49 first-in-man studies in healthy volunteers which randomised between IMP and placebo (or control), medians were as follows: six cohorts (mean 5.3, standard deviation (sd) = 2.4); 48 healthy volunteers in total, of whom 36 received IMP; and eight in the first cohort, of whom six received IMP. Thus, as in TGN1412 first-in-man study, allocation ratios strongly favoured IMP.

To be robustly informative, **we recommend** that the above 30 months' census of MHRA's 32 first-in-man studies of biological/biotechnology IMPs needs to be extended by international

regulatory collaboration; and likewise for approved first-in-man studies of chemical compound IMPs. In particular, the proneness of first-in-man studies of biologicals/biotechnologies to SUSARs in their initial cohort should be investigated.

**Table 2.3**: **Notifications to UK's competent authority (MHRA) about serious adverse events in 'first-in-man' Phase 1 studies in healthy volunteers 1 April 2005-31 March 2006**

| | |
|---|---|
| Number of 'first-in-man' *studies* which notified *any* serious adverse events in the notification period | At least 1 |
| Number of *subjects* about whom notifications were received (in the notification period) of *any* serious adverse events which had occurred in 'first-in-man' studies | 6 |
| *Maximum number of notifications* received per 'first-in-man' study for which *any* notification of serious adverse event was received (in the notification period) | 6 |
| *Maximum notification rate for a new experimental medicine* per 'first-in-man study for which any notification of serious adverse event was received for a new experimental medicine for a) initial cohort,<br>b) any dose-escalated cohort,<br>c) final dose cohort | a) 6/6<br>b) not done<br>c) not done |
| Number of 'first-in-man' *studies in which a new biological/ biotechnology was being tested* and which notified *any* serious adverse events in the notification period | At least 1 |
| For each notified study summary of its IMP (chemical versus biological/biotechnology) and its key design features | Initial cohort (out of 4 cohorts) of 8 healthy volunteers randomised 6: 2 to intravenous anti-CD28 monoclonal antibody (target = T cells) TGN1412: placebo with all eight administrations completed within 2 hours at protocol-unspecified 10-minutely intervals. Investigator Brochure stated that: "subjects will be closely monitored for Cytokine Release Syndrome" – but did not specify *how*. Of those randomised to TGN1412, 6/6 suffered CRS |

## *2.4. European Competent Authorities' regulation of Phase 1 studies, including 'first-in-man'.*

*Harmonised applications:* Applications for protocol approval are both harmonised across drug regulatory authorities in the European Union and designed to facilitate submission also to an independent research ethics committee (REC). Submissions for protocol and ethical approval could be sequential, or simultaneous. As of 1 May 2006, however, protocol-approval by MHRA must precede submission for ethical approval in UK of a 'first-in-man' Phase 1 study which involves administration of an IMP that targets the human immune system. The sponsor of each such study gives a signed undertaking to inform the drug regulator when his/her protocol-approved and ethically-approved 'first-in-man' study actually takes place, and the results thereof.

Since May 2004, MHRA has had the responsibility of approving all protocols for Phase 1 studies, including 'first-in-man' studies, in the UK. Prior to May 2004, ethical approval was required but there was no formal requirement on protocol approval (with some exceptions).

Currently, each protocol-approved study requires ethical-approval by a properly constituted independent local (or multi-centre) research ethics committee (LREC or MREC). In UK, there is ideally a statistician-member on such research ethics committees (Williamson et al., 2000), and statistician-membership of MRECs is usual. Research ethics committees can be specifically recognised, or designated, for their competence in dealing with different types of study, such as BMEC's recognition for healthy volunteer studies. RECs may take it that protocol-approval by UK's competent drug regulatory authority (MHRA) is effectively peer-review of the underlying science and experimental design.

*Advantage of harmonised applications:* A signal advantage of European drug regulatory authorities' harmonised applications for protocol-approval 'to commence a clinical trial in early phase of development' is that the same key data about the designs of protocol-approved 'first-in-man' versus other Phase 1 studies can be readily documented, and reported on annually, by each Competent Authority for:

a) all protocol-approvals which related to biologicals or biotechnologies;
b) 1 in 10 random sample of protocol-approvals which related to chemical compounds (with or without marketing authorisation).

**We recommend that:** European drug regulatory authorities agree on, and make public, an analysis protocol and publication-schedule (as for National Statistics) according to which they will report annually on the Phase 1 study designs that gained protocol-approval. Analyses should deal appropriately with four salient cross-classifications: a) 'first-in-man' or not, b) biological/biotechnology or not, c) healthy volunteers or not, d) serious adverse event reported, or not.

### *2.5. European drug regulatory authorities' example protocols for Phase 1 studies*

European drug regulatory authorities' harmonised application is accompanied on the website of Medicines Control Agency Clinical Trials Unit (now MHRA) by Annex 5 which gives illustrative guidance in mock applications for two fictitious products:

a) **chemical compound AB1234** - a proton pump inhibitor being developed for the treatment of gastro-oesophageal reflux disease, and

b) **monoclonal antibody MF6387** - a chimeric (human-murine) antibody against XY12, for which in vitro studies have been conducted to characterise affinity, specificity and biological activity. (Mechanism of action is due to MF6387's inhibition of the binding of XY12 to its receptors. XY12 is a growth factor for a number of solid tumours. Anti-angiogenic activity of MF6387 will also contribute to its inhibitory effect on tumour progression and promote tumour regression.)

The mock applications have serious statistical lacunae.

Quite basic statistical reporting standards (Altman et al., 1983) are transgressed throughout, including in summary of the kinetic data for oral AB1234 in animal species, where the numbers studied and mean (sd) per species for each summary statistic ($C_{max}$, $T_{max}$, $AUC_{0-24}$, $T_{1/2}$) were lacking. In AB1234's pharmacology section, the injunction appears: "Extensive information on study design not needed".

Statistical reporting standards are similarly inadequate for Toxicology. What objective interpretation can be placed on 'maximum non-lethal dose of 800 mg/kg in mouse versus 850 mg/kg in rat' without assessors' knowing either standard error of the estimate or even how many mice (rats) were exposed per single dose or at which doses in the range 200 to 2,000? Safety pharmacology studies are described as having been completed in accordance with the relevant ICH/CPMP Note for Guidance (International Conference on Harmonisation, 2000), which apparently gives licence for the assessor to know *only* the function tested (eg renal), the species tested (dog), the dose and route of administration (10mg/kg intravenously) *and* for essentially quantitative results to be summarised qualitatively (transient decrease in blood flow). See chapter 3.

Similar statistical reservations apply in respect of non-clinical pharmacology and Toxicology data presented for MF6387 (see, for example, the reporting of affinity which is unsupported by any measure of variation despite, or because of, its derivation from "BIA evaluation software"). There was a failure to document numbers of tumour-bearing mice in survival studies.

By contrast the clinical data section, where the proposed clinical trial is described, is clearly written for AB1234. A single dose, open label (that is to say not blinded) , dose escalation pharmacokinetic/pharmacodynamic study in young (18-45 years), healthy male volunteers is proposed, in which the starting dose of x mg is expected to have been suitably justified and likewise the step-wise increments of y mg. Cohort size is three volunteers throughout, that is for initial and incremented doses, and there are (expected to be) pre-defined rules for ceasing the dose incrementation (which, of course, protect from harm the next cohort of three volunteers). Maximum number of volunteers who risk harm is thus three per cohort, and the open-label nature of the study means that this is wholly transparent to the healthy volunteers. The study design does not, however, specify whether volunteers are randomised to cohort, which would be good practice.

The clinical data section for MF6387 ruled out studies in healthy volunteers in view of MF6387's potential anti-angiogenic activity and possible effects on the nervous system. MF6387 had not been previously administered to man, but SC1111, a murine antibody against XY12, had been used in a multiple dose Phase 1 study of 10 patients with solid tumours (summary of design and results provided). Expression of XY12 in corneal epithelium and in the axons of the central and peripheral nervous system had been reported in the literature. Cross-reactivity studies with normal human tissue had also localised MF6387's activity to axons in the brain, spinal cord and peripheral nerves.

The murine antibody had a median circulating half-life of 6 days, which was why weekly administration was proposed for MF6387. An informal argument was offered by which SC1111's doses by daily intravenous infusion over 14 days were translated to MF6387 doses to be given subcutaneously and weekly. A further check was made that the maximum thus-translated MF6387 dose (200mg) was also at least 10-fold less than MF6387's maximum pre-clinical dose of 50mg/kg. For this check, minimum patient weight of 50kg was assumed and gave the maximum proposed human dose as 4mg/kg.

The MF6387 application described its Phase 1 study design as randomised, placebo controlled, double-blind, parallel group with 90 adult patients with clinical diagnosis of solid tumour, measurable or evaluable disease, and life expectancy of at least 6 months being recruited at multiple sites in multiple member states. Amongst its precautions, the protocol signalled that all patients would be monitored for the development of antibodies to MF6387, and would undergo

both slit lamp examination to ensure no adverse corneal effects and neurological assessments due to the observed cross-reactivity with central and peripheral nervous system tissue. Inter-patient administration interval was not formally specified but patients (unlike healthy volunteers) usually enter studies serially rather than being convened on the same day. Withdrawal criteria included an adverse event during or within an hour of administration.

The protocol warned that all data remained the property of Drugs-R-Us. No publication without the consent of Drugs-R-Us would be allowed.

*Key statistical lacunae in mock applications were:*

1. Number of animals per species, and designs for pre-clinical animal studies omitted.
2. Qualitative words used to describe quantitative data when reporting the results of pre-clinical studies: appropriate measures of statistical variation were lacking.
3. Rationale for why healthy volunteers, rather than patients, were recruited for 'first-in-man' study of AB1234 was lacking. (MF6387's protocol and investigator brochure, on the other hand, explained several grounds for the inappropriateness of its recruiting healthy volunteers.)
4. Omission of a specific section which summarises a priori risk, including how exceptional or novel the Investigational Medicinal Product is.
5. Omission of inter-administration interval between successive human volunteers in the first cohort of participants within a 'first-in-man' study. (Considerations may include mean and variability of time to maximum concentration, estimated half-life in pre-clinical studies, and plausible time of onset for anticipated, but unlikely, harms.)
6. Omission of time delays between the taking of biological samples for safety monitoring, and the availability of results based on those samples.
7. Omission of re-imbursement scheme for healthy volunteers means that competent authority is not in a position to monitor re-imbursement against volunteer's time, inconvenience, or a priori risk.
8. No required reporting of 'first-in-man' study results to competent authority within *pre-defined* number of weeks of effective commencement of the trial (to be specified by competent authority on study-specific basis).

*TGN1412's MHRA-approved 'first-in-man' Phase 1 study protocol appeared to meet expected standards as conveyed by mock applications to the Competent Authority.*

### *2.6. Assessors' reports to Competent Authority on TGN1412 protocol, and information sheet for volunteers*

*Assessors' reports:* Three assessment reports were made. The first relates to quality (wherein TGN1412 was clearly described as 'a new presentation of a new drug substance' and as 'recombinant humanised agonistic anti CD28 monoclonal antibody'). Secondly, pharmaco-toxicological (safety) considerations are made (wherein TGN1412's extreme novelty was well described in the assessor's second paragraph on Pharmacology: 'TeGenero's novel TGN1412 bypasses the requirement for T cell receptor (TCR) triggering and activates T-cells irrespective of their TCR specificity. TGN1412 therefore represents the first universal T-cell growth factor applicable for therapeutic purposes in the intact organism'). The third concerns medical aspects.

The pharmaco-toxicological assessor paid particular heed to the paradox of cross-reactivity with central nervous system (CNS) tissue yet no CNS-related observations in toxicological or histological studies in cynomolgus monkeys; and to rats' JJ316 mediated T-cell expansion and activation appearing to be faster than in non-human primates treated with either TGN1112 or

TGN1412 and associated lower estimated half-life in rats. S/he noted that systemic exposure to TGN1412 increased by up to 20-fold as dose increased from 5 to 50 mg/kg, and that there was evidence for increased mean terminal half-life of TGN1412 as dose increased. In the toxicology section, the assessor noted that TGN1412 had significantly lower pharmacological activity in rhesus monkeys compared to TGN1112 (one and two rhesus monkeys respectively). Presciently, the assessor suggested that this may be explained by different affinities or FcR binding properties of the two antibody formats. Cynomolgus monkey was selected as the more appropriate non-human primate species – in which the immunogenicity of TGN1412 was described as 'low', since only four out of 16 treated animal showed substantial titres of anti-TGN1412 antibodies in serum at 3 to 4 weeks after dosing.

The medical assessor's summary of the design of the proposed healthy volunteer study followed, rather uncritically, the submitted protocol and questioned neither omission of inter-individual dosing interval, the specification of clinical interventions for some anticipated risks, nor the composition of a claimed independent Data Safety Monitoring Board. Despite that follow-up should be for seven weeks in view of the risk of anti-TGN1412 antibodies, with maximum tolerated dose dependent on TGN1412's effects on the human immune system, and that volunteers would be closely monitored for lymph node and spleen enlargement, the medical assessor summarised the risk: benefit ratio in *healthy volunteers* as favourable. CRS risk was not specifically addressed.

*Information sheet:* The information sheet for healthy volunteers in TGN1412 Phase 1 study explained that the Parexel Clinical Pharmacology Unit (PCPU) was based at Northwick Park Hospital (but is independent of the hospital), and that Parexel was being paid by TeGenero to conduct a 'First Time in Human' study of TeGenero's new compound TGN1412, a monoclonal antibody that targeted T cells, which was being developed for the potential treatment of various inflammatory diseases, such as rheumatoid arthritis, and possibly also for a type of leukaemia.

As far as we are aware, volunteers were not told that allocation ratio was 6:2 (TGN1412: placebo) but were advised that: "at the allocated time (approximately between 08.00 and 10.00), you will be given your study medication in the form of an intravenous infusion (where the study drug is administered directly into a vein). For practical reasons, the time of dosing will be staggered . . . "

Possible side effects were listed, but *less explicitly in Version 02 Final (09 February 2006)* – for example, the terms immunosuppression, autoimmunity and cytokine release were deleted - than in Version 01 Final (23 December 2005). (Both versions are available from BMEC, which had commented that the information sheet was too technical and needed to be simplified.). Volunteers were warned that: "this study may involve risks that are currently unforeseen. No significant side effects have been seen in the animal studies, and although these are not a precise indicator of what will happen in humans, they give some indications of the possible side effects".

A further paragraph explained that risk of anaphylaxis applies to all studies at PCPU, with drugs at every stage of development, and that the staff were well trained in anticipation of this (unlikely) possibility.

How well-equipped both PCPU and the intensive care unit at Northwick Park were to cope with cytokine release, and particularly with CRS, was not addressed.

*The information sheet for healthy volunteers did not*:

a) Explain that the investigators believed that TGN1412 selectively activated T-regulatory cells (see chapter 3), which was the rationale for its use in autoimmune disease. But, *if selectivity failed to occur in humans*, then trouble from T-cell over-stimulation could be expected
b) Explain that an anti-CD3 monoclonal antibody with Fc receptor binding ability - which TGN1412 also has (but to a lesser extent) – *had previously caused cytokine release storm.*
c) *List Cytokine Release Storm (CRS) – by name - as an anticipated Possible Side Effect* on page 6 of Version 02 Final (09 February 2006) of Patient Information Sheet.
d) Reveal that *randomisation ratio was 6: 2* for TGN1412: placebo in the lowest intravenous dose group, 0.1mg per kg of body weight.
e) *Explain why healthy volunteers* rather than patients with inflammatory diseases were recruited for TGN1412's "first-in-man" study.
f) *Identify which independent Research Ethics Committee had approved* the TGN1412 "first-in-man" study.
g) Explain whether volunteers would be given a *copy of the approved study-protocol on request*, and which details (if any) were commercial-in-confidence and would be blanked-out from the released protocol.

### *2.7. Risk classification and recompense for healthy volunteers*

The altruism of healthy volunteers, and of patients, who take part in 'first-in-man' studies of vaccines (such as against HIV or avian flu), new surgical procedures, or new experimental medicines is rightly admired.

The Textbook of Pharmaceutical Medicine (Posner, 2005) advises that volunteers should be paid according to inconvenience rather than a priori risk, partly because - if harm does befall – volunteers will be properly compensated. (TeGenero has, however, declared itself bankrupt in the aftermath of TGN1412 Phase 1 study in healthy volunteers.) In general, riskiness is dose-related, but also depends on the nature and novelty of the Investigational Medicinal Product, and on the *uncertainty* of predictions from pre-clinical studies.

Re-imbursements for healthy volunteers are approved, in effect, by a range of independent RECs, part of whose remit is to review the information sheet for healthy volunteers. In effect, there is no centrally-held information on re-imbursements for volunteers that can be readily monitored, analysed in accordance with measures of inconvenience and a priori riskiness, and annually reported upon in a public forum.

Higher per diem re-imbursement according to a priori risk would, of itself, signal to healthy volunteers the degree of novelty or biological uncertainty that was associated with a new experimental medicine to be tested on them.

**We**, and others (Kenter & Cohen, 2006), **make recommendations** on a priori risk classification, and we propose public debate about whether there should be risk-calibrated re-imbursement of healthy volunteers. Even modest rewards can themselves jeopardise safety if they induce volunteers to conceal pertinent risk factors.

### *2.8. Syntheses by drug regulatory authorities*

Since May 2004, UK's drug regulatory authority, together with those of other European nations, has been well placed to make a publishable synthesis of the designs, analysis and reporting

standards, and outcomes from the pre-clinical studies that are reported in applications made to a Competent Authority for protocol-approval of a 'first-in-man' Phase 1 study of an IMP. Separate syntheses should be considered for chemical compounds, monoclonal antibodies, and other biologicals/biotechnologies.

Planned synthesis across a range of studies imposes scientific discipline, in terms of written protocols to determine eligibility criteria, confidentiality for sponsors, and the data to be abstracted from individual submissions, so that rigorous synthesis can be achieved. The existence of such protocols helps to ensure that reporting in the individual submissions is adequate by well recognised statistical standards(International Conference on Harmonisation, 1999). Assessors would otherwise have to go back to sponsors to obtain the missing details on design, numbers of animals or tissues studied, and standard deviation of key outcomes from an important range of pharmacological and toxicological pre-clinical studies that are the essential backdrop to subsequent clinical studies.

**Moreover, we recommend that** Europe's drug regulatory authorities could, and should, properly document, on a periodic basis: a) the experimental designs (including allocation ratios, cohort sizes per dose, doses and inter-administration waiting times) for 'first-in-man' studies to which they give protocol approval, and b) the extent to which sponsor-anticipated risks materialised, how unlikely those harms proved to be which were specifically identified on theoretical or other grounds but judged a priori by the sponsor as unlikely to occur; and the risk to volunteers of unprecedented harms.

Separate syntheses of study designs and of the occurrences of predicted, theoretical and unprecedented harms - either as adverse events or distributional changes in biomarkers – should be considered for healthy volunteers and for patients, and by type and novelty of compound, and by a priori risk stratum.

Over time, all concerned – from sponsors through assessors to volunteers and the public – should gain a better understanding of the intellectual nature, complex decision-making and fine judgement in the balancing of risks and benefits that are entailed in the early development of drugs.

The last of the proposed syntheses - on different sorts of harms – cannot be conducted, even by drug regulatory authorities, unless there is a legal obligation on sponsors to pre-notify the competent regulatory authority of the actual start date of a protocol-approved 'first-in-man' study and, likewise, an obligation to provide the competent regulatory authority with, minimally, a set of *pre-determined* data summaries within W1 weeks of the study's pre-notified start date.

As Goodyear remarked (Goodyear, 2006), the aftermath of TGN1412 should foster a more open culture that is rewarding and safeguarding of all concerned. Some a priori risk assessments will turn out to be wrong, that were nonetheless made in good faith and best judgement. That is the nature of uncertainty, but we can develop methods for quantifying better how frequently, or infrequently, our summary a priori assessments are unduly optimistic and harm has come to healthy volunteers.

New biologicals and biotechnologies, present new challenges to pharmaceutical industry, professionals, volunteers and the public. Statistical science can help to meet that challenge through better reported pre-clinical study designs and outcomes, and by appropriate and skilful synthesis of evidence from different sources.

# 3. Immunological Considerations

This special chapter contributed by Sir Peter Lachmann, FRS, FMedSci considers some particular issues to do with the immunology of TGN1412.

## *3.1. Properties of the antibody*

TGN1412 is a humanised monoclonal antibody of IgG4 kappa type reacting with the T-cell co-stimulatory molecule CD28. This antibody, and the mouse monoclonal from which it was derived, has the unusual property of being able, on its own, to activate CD28 positive T-cells– i.e. without also needing ligation of the T-cell receptor. For this reason it is known as a superagonist antibody and can activate whole populations of CD28 positive T-cells. More conventional anti-CD28 antibodies are also agonist in that they can substitute for the ligation of CD28 by its normal ligands, B7-1 and B7-2; but these antibodies give rise to activation of T-cells only when the T-cell receptor is also ligated and therefore cause activation only of a set of T-cells with a particular antigenic specificity.

Mapping of the antigenic site on CD28 reacting with TG1412 shows that the superagonist antibody reacts with an unusual epitope in the C"D-loop of CD28 which is different from the site bound by the conventional anti-CD28 antibodies or by B7-1 and B7-2. It is not obvious why this specificity alone should confer superagonist activity and two mechanisms have been proposed in explanation. The Wurzberg group who developed the antibodies suggest that the explanation lies in the unusual ability of these anti-CD28 antibodies to react bivalently with CD28 and therefore presumably to cross-link neighbouring CD28 molecules(Dennehy et al., 2006). Such cross-linking can be envisaged to enhance signalling so as to produce activation. A further suggestion has been made by Colaco (Colaco, 2006 ) that, by analogy with the widespread T-cell activation produced by the OKT3 monoclonal antibody, (which reacts with a component of the T-cell receptor), that the explanation lies in co-ligation of an Fc receptor on the cell surface by the Fc portion of the monoclonal antibody. This phenomenon accounted for the T-cell stimulation by OKT3 antibodies that could bind Fc receptors; and it was shown that engineered OKT3 antibodies that were unable to react with the Fc receptors were immunosuppressive. It would not be difficult to test whether either or both of these mechanisms are operative by testing Fab and Fab'2 fragments of the antibody in-vitro.[1]

Of greater importance to what occurred in the trial, TGN1412 is claimed to have the further unanticipated property of, in-vivo, selectively activating regulatory T-cells, a subset of CD4 positive T-cells that also carry CD25 on their membranes. This is the basis for developing this antibody as a therapeutic for autoimmune disease (in addition to using it for expanding T-cell numbers in haematological malignancies) and there are data from studies in rodents that it has a beneficial effect in models of autoimmune disease. Nevertheless this selectivity for regulatory T-cells is surprising since CD28 occurs on nearly all CD4 positive T-cells (and about half of CD8 positive T-cells). In one of their publications (Beyersdorf et al., 2006) the Wurzburg group do propose a mechanism for this selectivity. They propose that there is initial stimulation of all CD28 positive T-cells which causes "conventional" T-cells to secrete Il-2. This Il-2 then reacts with CD25 (which is an IL-2 receptor) on the regulatory T-cells which, together with the

[1] New (November 2006) results reported from NIBSC show that only immobilised antibody has superagonist activity in-vitro. This supports the idea that cross-linking to a receptor on a near-by cell is the basis of the superagonist activity. In the microenvironment of a lymph node, dendritic cells and macrophages are in close proximity to T cells and these cells express Fc receptors. It is also possible that complement fixation and reaction with a complement receptor could be involved.

anti-CD28, gives them a powerful activation and proliferation stimulus and enables them to control the stimulated "conventional" T-cells. Assuming this explanation to be correct, it should perhaps have suggested some caution in the use of the antibody. Such a mechanism would presumably be very dependent on the balance between the number of "conventional" CD4 positive cells and the number of CD25 positive cells so that the amount of Il-2 produced is sufficient to stimulate the regulatory cells but not so large as to cause toxic reactions, and that the number of regulatory T cells is sufficient then to control the activity of other stimulated T-cells. It may well be the case that this balance can be achieved more easily in one species than another; and in a single species, such as humans, it may be affected by age, sex, clinical status, the presence of infection or of an acute phase state… or other factors. In the TGN1412 trial the antibody was given to healthy young men between the ages of 18 and 40 in whom the balance needed for achieving the selective stimulation of regulatory T-cells was clearly not in place. It is worth pointing out that adults of this age are believed to be particularly prone to T-cell hyper-stimulation and this is blamed for the observation that so many people of this age group died of what appears to have been a cytokine storm in the 1918 'flu epidemic .

TGN1412 was tested both in rhesus monkeys and in cynomologous monkeys. CD28 from the latter species appeared to have the stronger reaction with the antibody. In-vivo dosing of cynomologous monkeys did show that plasma levels of two cytokines, Il-6 and Il-5, were quite markedly increased by treatment with the antibody (approximately 9-fold at 5mg/kg and 18-fold at 50mg/kg - with wide ranges of the individual values) while two important pro-inflammatory cytokines, TNF$\alpha$ and interferon gamma, showed no increase. There was a transient increase in T-cell numbers (both CD4 and CD8 positive) The investigators did not regard the findings as having major significance and the animals showed no signs of illness at any of the doses tested up to 50mg/kg.

### *3.2. The Northwick Park "first-in-man" trial*

Six male adults aged between 18 and 40 and in good health were given TGN1412 intravenously at a dose of 0.1 mg/kg and all developed the typical manifestations of a "cytokine storm".

In view of this unexpected and exaggerated reaction in man at what appears to be a much lower dose of the antibody than used in the monkeys it is relevant to ask whether the equivalent doses were appropriately calculated. For an antibody, effective dose is the product of concentration and avidity. In this study concentration only was used and I have been unable to find measurements of the avidity (or affinity[2]) of TGN1412 for either cynomologous monkey CD28 or human CD28. Avidity measurements are readily done in-vitro and should be available whenever doses are calculated on the basis of animal data using cross reaction with the antibody to be used in man. Sequencing the epitope in both species and demonstrating identity is not an adequate substitute for measuring avidity. Antibodies recognise conformation and this can be affected by mutation elsewhere in the molecule. A good example is a monoclonal antibody that recognises only one of two allelic forms of C7 (a complement component). These differ by a single amino acid substitution which is 40 amino acids n-terminal to the epitope region (Wurzner et al., 1995).

For obtaining an equivalent dose for use in-vivo it is also possible that body weight is not the best denominator. A case could be made for using an estimate of total extra-cellular fluid volume

[2] Affinity measures the strength of reaction of a single antibody binding site with its antigen; avidity measures the strength of the reaction between the whole antibody and the antigen (in this case CD28 on the cell surface). For a Fab fragment they have the same value. For an intact IgG molecule reacting bivalently the avidity is much higher than the affinity

or of total T-cell number instead. However, it is not likely that correction of the effective dose for avidity and using one of the alternative denominators would make sufficient difference to the calculation of equivalent dose to account for the toxic effect being produced by just 0.1mg/kg in man while the monkeys tolerated 50mg/kg. The human volunteers do seem to have been more susceptible than the monkeys.

In the Investigator's Brochure the possibility of a cytokine storm, by analogy with the OKT3 experience, was envisaged, and it was said that "*appropriate counter-measures must be taken*". What these measures should be was not set out and, in the event, it appears that they were not immediately in place. The Northwick Park Hospital intensive care team have now published an informative account of the management of the six patients (Suntharalingam, et al., 2006) . Diagnosing and dealing with cytokine storms is difficult and requires a multidisciplinary approach including experienced support from clinical immunology, as well as the relevant organ specialties - not all of which seem to have been initially available. From the account given it would appear that the diagnosis of cytokine storm was not made for some hours after the onset of symptoms.

Furthermore the facilities to perform some of the more speculative, but probably effective treatment approaches such as plasma exchange (to remove a proportion of the antibody) or acute dialysis using high molecular weight cut-out dialysis membranes (to remove some cytokines) do not appear to have been available either.

An antagonist antibody to the IL-2 receptor CD25 (dacluzimab) was given for three days starting about 24 hours after the infusion. Since it is the ligation of CD25 to which the investigators ascribe the selective activation of regulatory T cells there could be some doubt whether this was the cytokine inhibitor of first choice. No cytokine inhibitors to counteract TNF alpha (Infliximab or Enbril) or IL-1 (Anakinra) were given. The cytokine assays presented show a dramatic, very early rise in TNFalpha suggesting that administration of a TNF alpha antagonist immediately after the onset of symptoms might have been beneficial.

It may be legitimate to ask whether studies, where even the remote possibility of complications such as cytokine storms has been foreseen, would not better be carried out only in hospital settings with full facilities for giving the appropriate tertiary care without delay.

Finally it should be pointed out that the cytokine storm seen is not a side effect in the true sense of the word – it is a consequence of T-cell activation - the primary effect produced by the antibody. What seems to have gone wrong is that in this target population and with the dose of antibody used the T-cell activation did not achieve the selectivity for regulatory T-cells which has been observed in animals. It is possible, but not certain, that doing more extensive in-vitro studies with mixtures of different numbers of T-cell sub-sets and of dendritic cells and macrophages; and using various concentrations of antibody to determine the precise conditions needed for achieving regulatory cell stimulation as opposed to pan T-cell stimulation might have given an indication that this might happen. Furthermore, used under more closely defined conditions and for appropriately selected indications, TGN1412 could still prove to be of substantial clinical benefit. It would be a pity if all development were abandoned as a result of this initial trial.

# 4. Ethical Issues

In this section we discuss ethical issues from a statistical perspective, including assessment and communication of risk, and draw distinctions between risks at the level of the individual and society.

## *4.1. Introduction*

It is clear that 'first-in-man studies' raise a number of ethical issues. This is true even where, as in the case of cytotoxic studies, such trials are run in patients, whom it might be hoped would benefit from the treatments being given (even if that hope is fairly slight). Amongst standard ethical concerns, the most contentious, where the subjects are healthy volunteers, is that of *non-maleficence* (the obligation not to inflict harm intentionally), since no compensating benefit in terms of health can possibly result from the treatments given and indeed, as TGN1412 showed, harm may result.

There is a considerable statistical literature, much of it based on decision analysis, to do with ethics and risk in clinical trials. For example, one might cite the extensive work on the continuous reassessment approach to dose finding in cancer initiated by O'Quigley and co-workers (O'Quigley, Pepe, & Fisher, 1990), that on so-called 'bandit designs' by Berry and colleagues (Berry & Eick, 1995) and Bayesian approaches to randomisation and equipoise as outlined in the book edited by Kadane (Kadane, 1996). Others have looked at ethical issues raised by research in and treatment of vulnerable populations, such as children or patients in developing countries (Garrett, 2006; Hutton, 2000), problems associated with the use of placebos in randomised trials (Senn, 1997a; Senn, 2002) , approaches to obtaining informed consent (Zelen, 1992), approaches to randomisation (Gore, 1994; Rosenberger & Lachin, 2002), alternatives to conventional randomisation (Trochim & Cappelleri, 1992) or even ethical standards and guidelines governing the work of the statistician (Hutton, 2000).

Ideally the report of a Royal Statistical Society working party should be in a position to give some firm recommendations, perhaps based on decision analysis, regarding acceptable risks and their assessment in first-in-man studies. Unfortunately, we are not in a position to do this. First we think that the problem is a difficult one that is urgently in need of further study; secondly we think that a much wider debate is needed. First-in-man studies in healthy volunteers raise a number of extremely unpleasant ethical problems that can be solved only at a societal level. Instead we raise some issues and make tentative and incomplete recommendations below.

## *4.2. Issues*

### *4.2.1. Informed consent*

#### *4.2.1.1 Utmost good faith*

It is obviously impossible for a healthy volunteer to enter fairly into a contract to provide his or her body for the purposes of medical research if the other contracting party withholds relevant information. The issue is analogous to that of utmost good faith (Faunce & Bolsin, 2005; Wilkie, 1997) as applied in the field of insurance. The standard to be applied by trialists and sponsors should be that of the 'open protocol', which is to say that, amongst other matters, whereas the result of the allocation of treatments itself may be hidden from the subject, to the extent that those results are governed by randomisation, all other aspects of the allocation procedure must be revealed. Although, when dealing with patients, there are some practical limits to informed consent that may have to be observed, nevertheless the spirit in which consent is obtained should be one of empowering the subject to make and if necessary rescind decisions (Ioannidis, et al.,

2004). This would seem to apply *a fortiori* for healthy volunteers, since there should be no medical circumstances that complicates their decision.

As an example of 'open protocol, hidden allocation', consider a trial with three doses, D1, D2, D3 and a placebo and where allocation is to proceed in three steps with (say) six subjects being randomised (three to D1 and three to placebo) in step one and then a further six (three to D2 and three to placebo) in step two and then a final six (three to D3 and three to placebo). To observe the standard, this general scheme must be revealed to the healthy volunteers concerned, even though, at any dose step they will be unaware as to whether they will receive the active treatment or placebo. Thus, the subjects become aware which dose is *not* given by knowing the dose step to which they are contributing, whilst remaining unaware as to whether placebo or active treatment *is* given. (Such trials in which partial information as to allocation is or ought to be available have been referred to as *veiled* (Senn, 1995; Senn, 1997b),are frequently encountered in connection with first-in-man studies and carry with them implications as to analysis, which are considered in the chapter on design.)

It will also be necessary for the healthy volunteer to be provided with an assessment of estimated risks of serious side effects. Also, if the trial is to proceed until (presumably mild) tolerability problems are encountered, this should be made clear.

#### *4.2.1.2 Competent risk assessment*

The standard of utmost good faith, however, is clearly not enough. It would be quite possible for a sponsor to share everything known with a volunteer simply because not much was known and very little effort had been made to discover it.

There are some technical issues here. From the point of view of an immediate decision, a probability of a probability is still a probability. Thus, to take a trivial and unrealistic example, if it were believed by a researcher (say Smith) with probability 0.99 that the probability of a side effect was 0.001 and with probability 0.01 that the probability was 0.8 (because say, fairly confident that the drug was relatively safe but with a slight nagging doubt that it might be very risky) then, according to the belief of that researcher, the overall probability of a side effect in any *single* subject about to be given treatment would be $0.99\times 0.001+0.01\times 0.8\approx 0.009$. As regards that researcher's assessment of the probability that the first healthy volunteer will have an adverse reaction, this is no different from a researcher (say Jones) who believes with absolute certainty that the probability is 0.009.

However, Smith and Jones, would differ in their probabilities that, if *two* individuals were treated, *both* would show an adverse reaction. Jones can argue that since he is absolutely certain that the probability of a side effect is 0.009 in any given subject, observing a subject with one side effect does not change this belief. Hence the probability of a second side effect *given* the first is still 0.009 and the probability of two side effects can be obtained as 0.009 × 0.009. For Smith, the evidence of one side effect increases the belief that the drug is of the more dangerous sort. In fact, applying Bayes theorem, we would have that the probability of the drug being very risky given that a side effect had occurred would be the joint probability of this event $0.01\times 0.8=0.008$ rescaled by dividing by the probability of a side effect occurring at all, namely 0.009. Hence the posterior probability of the drug being very risky becomes $0.008/0.009\approx 0.89$ and of being less risky becomes $1-0.89=0.11$. Carrying through to the prediction of a second side effect, this now becomes $0.11\times 0.001+0.89\times 0.8\approx 0.71$. Hence, for Smith, unlike Jones, the second side effect is much more likely than the first.

These two cases are very different as regards the value of further information. In the case of Smith it may be argued that perhaps further background research would enable one to refine this uncertainty further before studying the drug in humans and that, if such refinement were possible without using human subjects, there was a moral obligation to undertake it, whereas in the case of Jones one might argue that nothing further may be learned by studying the drug in humans.

Although it would be unreasonable to expect that all uncertainty about risk be eliminated before undertaking research in humans, since the further reduction of such uncertainty is the point of such research and the fallibility of biological theory makes it necessary, it should also be established that due care has been taken in both background experimentation and calculation to quantify risk and uncertainty about risk.

#### *4.2.1.3. Adequate communication of risk*

As is generally recognised, communicating risk is a notoriously difficult topic yet if subjects are to give truly informed consent in entering a clinical trial it is not sufficient that the risks are understood by the persons running the trial, it is also necessary that they are understood by the subjects entering the trial. Thus good communication of risk is essential to the ethical conduct of trials.

The Royal Statistical Society has a long tradition of being concerned not just with all the technical aspects of assessing and measuring risk but with the business of communicating risk clearly. In his presidential address to the Society, Adrian Smith (Smith, 1996) covered the communication as one of a number of matters that needed to be improved in Society's handling of risk. In 2003 the Royal Statistical Society devoted a whole issue of its journal *Statistics in Society*, also known as *Series A*, to this topic and in the same year a special booklet on this topic was issued by the Society (Watts, 2003). In that booklet important points that we endorse were made as follows.

First, that the communication of risk should be made as concrete as possible. For instance in the 2003 issue of *Statistics in Society* referred to above (Simpson & Lee, 2003) report Richard Peto's graphic description (Peto, 1994) of the risks attending 1000 twenty-year-old smokers if they do not give up as follows

> on average, about one will be murdered and six will die in motor vehicle accidents (often perceived as the greatest risk of premature death in the UK), but about 250 will be killed by smoking in middle age (before 65 years of age)….

Second, that despite the fact that simplicity is a virtue, risks should be expressed numerically, at the very least to the extent of showing some difference in degree.

### *4.2.2. Acceptable Risk*

#### *4.2.2.1 Determining the acceptable risk*

As London points out (London, 2006), the fact that the highest standards of informed consent are observed does not excuse trialists from the moral obligation of assessing risk and determining if it is acceptable. He proposes that notions of acceptable risk should be guided by the levels of risk that individuals voluntarily accept in other spheres, for example in leisure pursuits such as rock-climbing or everyday activities such as driving.

It is here that there is a need for an extensive debate, involving not just scientists carrying out clinical trials, but also wider society. It is pointless to deny that there is some risk attendant on medical progress. In this respect views commonly expressed on a variety of medical issues, and sometimes by the same persons, are quite inconsistent. For example, the National Institute for

Health and Clinical Excellence has been criticised for not showing greater haste in recommending the use of Herceptin more widely in women suffering from breast cancer by some of the same newspapers that have criticised indecent haste in clinical research as a result of the events attending the TGN1412 trial.

#### *4.2.2.2. Dose justification*

By starting with minute doses and incrementing in minute steps it would be possible to reduce to almost zero the risk of an adverse reaction at the expense of adding several years to existing drug development. We suspect that this is not, in fact, what 'society at large' would desire but it points to the fact that if progress in medicine is desired some element of risk is inevitable.

#### *4.2.2.3. The appropriate metric for acceptable risk*

We draw attention to a paradox in defining acceptable risk. Suppose that it were argued that an acceptable level would be a risk of no more than 1 in 10,000 for a serious side effect but that careful expert assessment for a new pharmaceutical placed this risk (initially) at 1 in 5,000 for a healthy volunteer taking the drug. It would appear that the drug cannot be studied. However, by recruiting two healthy volunteers, one to be randomised to placebo and one to drug, the risk for any one subject is halved to 1 in 10,000 and hence (just) acceptable. It is doubtful, however, that anybody would accept this as being a reasonable device to make the drug capable of being studied. It is interesting to note that if one considered either a) the premium that would be required to make the trial insurable or b) the expected number of serious side effects that would result from running such a trial, then neither of these would be reduced by increasing the number of placebo subjects in such a trial. Note also that, because of different expectations of benefit, what is an acceptable risk for an individual may be unacceptable to society and *vice versa*.

Therefore, we suggest that for the purposes of judging the acceptability of a *trial* (and possibly even a drug development programme*)*, the standard to observe is the expected number of serious side effects it would produce (taking due allowance of the fact that the design may permit early stopping if problems occur) whereas a given individual deciding whether to participate or not will properly wish to assess the risk to him or her, which is a different matter. In other words two important assessments of risk need to be made: first, on behalf of any participating healthy volunteer and second, on behalf of 'society'. Running the trial cautiously, for example by allowing early stopping may reduce the risk of the latter but this is of no relevance to a healthy volunteer deciding whether to take a drug or not. Similarly, satisfying the standard that a given healthy volunteer is taking an acceptable risk does not mean that the trial is acceptable in terms of expected number of side effects it might produce.

### *4.2.3. Adequate recompense*

There is a general repugnance amongst many commentators on the ethics of clinical trials to consider the issue of financial reward for participation. A common (implicit) point of view is that it is acceptable for everybody except patients to expect to make money from the development of new treatments. Patients, however, may volunteer for clinical trials either because it is their only opportunity to obtain a chance of a new treatment(Senn, 2001) or because of some desire to contribute towards finding some solution to a problem from which they suffer. With healthy volunteers, however, the issue of financial recompense is raised, and notwithstanding the common view that volunteers are recompensed for time and inconvenience and not risk, if our recommendation that explicit assessment of risks be provided to such volunteers is followed, it seems possible that a market will result, whereby riskier trials require higher compensation. Of course, there is even the possibility that some parts of a trial (cohorts of recruitment) may be regarded as carrying a bigger risk than others.

We do not feel that it is appropriate to make any recommendations regarding this but simply point out that this may be a logical consequence.

However, there is also the issue of compensation in the event of harm. Here we do consider that trialists ought to heed the warnings of Laurence, who has pointed out that many current trials in the United Kingdom use language in recruiting subjects, which violates standards of consumer law because implying compensation for subjects in the event of serious adverse reactions occurring without, however, making such compensation the subject of an official contract (Laurence, 2005, 2006).

For healthy volunteer studies, every subject should be insured and a formal risk assessment should be provided to the insurer. It is highly probable that insurers will also make an assessment of their own.

### *4.3. Conclusions and recommendations*

4.3.1. Before proceeding to a first-in-man study a formal assessment of risk should be produced which provides:
- A quantitative justification for the starting dose
- An estimate of risk for the recommended dose
- Estimates of uncertainty regarding these recommendations

4.3.2. These estimates should be available in a separate document to be provided to ethical review boards, subjects to be recruited into a trial and trial insurers

4.3.3. In addition a given first-in-man trial should provide:
- A justification of the proper interval between dosing subjects
- A justification of the dose steps the trial will use
- A general justification of the design including monitoring and stopping rules
- An estimate of the expected number of serious adverse reactions that the trial will produce

4.3.4. The standard of informed consent to be observed is ‘open protocol hidden allocation’ which is to say that all aspects of the trial design shall be shared with subjects to be recruited.

4.3.5. Risks should be communicated and expressed using numbers and not just words

4.3.6. All participants in a healthy volunteer study should be insured.

4.3.7. Debate and research is needed about maximum acceptable levels of risk for individuals and society.

# 5. Safety Databases

The quantification of risk of harm requires a process for the collection, dissemination and investigation of pertinent safety data. In this chapter, some previous attempts to collate and publish information on the safety of Phase I studies are reviewed and their limitations discussed. Current statutory requirements for the reporting of serious adverse event (SAE) data to regulatory agencies are then considered. Recommendations are made for broader access to a more comprehensive safety database to facilitate the quantitative assessment of risk.

## *5.1. Current data to assess risk of harm*

Sibille et al (Sibille et al., 1998) highlights that adverse events from Phase I studies are *rarely published, leading to a lack of information.* However, several researchers have at various times attempted to quantify the occurrence of serious-type adverse events in Phase I healthy volunteer studies, as reviewed below.

In Royle and Snell's report (Royle & Snell, 1986) on the Association of British Pharmaceutical Industry's request for data on healthy volunteer studies from its member companies in June 1984: Forty-three companies responded and - dependent up on their internal records - described their experiences over a period of between 1 and 6 years. The response rate is not reported, although the authors suggested that the 28 companies that responded who conducted at least some of their studies in-house represented the complete set of such companies in the UK. Adverse events were classified as death/life-threatening; serious but not life-threatening; and minor. Studies were separated into internal (that is, conducted at their own in-house unit) and external (that is, contracted out) – although the exact number of studies included in the survey was not reported. In terms of in-house studies, there were 18,671 subject entries (roughly 500 per year) with no deaths or life-threatening suspected events noted. Serious suspected reactions were reported in five (0.03%; 5/18671) subjects. For the external studies, there were 8,753 subject-entries (about 3,000 per year). There was one death (the well documented Cardiff medical student case) but no further life-threatening suspected events. (A well documented death at a unit in Ireland was not included, since neither the study nor sponsor company was UK based.) In addition, there were eight serious suspected events reported giving a crude percentage of 0.10% (9/8753).

Orme *et al* conducted a survey of members of the Clinical Section of the British Pharmacological Society with the aim of identifying all Phase I studies (together with all corresponding adverse events) conducted over a 12-month period from October 1986 to September 1987 (Orme et al., 1989). The survey was issued to academic, health service, pharmaceutical and contract research organizations and adverse events were classed as: severe (defined as potentially or actually life threatening or required hospitalization); moderate; and mild. The response rate was 87% and duplicate studies were removed from the database. In total, 8163 subject entries occurred in 1377 single dose studies and 375 multiple dose studies – with inevitably some subjects enrolled in more than one study. Three volunteers (0.04%; 3/8163) recorded a severe AE, with all subjects showing a full recovery. Broadly speaking, these three events were rash, anaphylactic shock and ulcer perforation and were considered to be so-called type B event (unpredictable and severe, not-related to dose, and possibly due to hypersensitivity or immunologic reactions) (Strom, 1995). By contrast, most of the 45 moderate events were considered to be so-called type A – that is, predictable, dose related, less serious and a result of the extension of the pharmacological effect.

More recently, Sibille *et al* described their experiences of conducting 54 Phase I healthy volunteer studies of 20 new chemical entities over a period of 10 years (1986-1995) at their in-house unit in France (Sibille, et al., 1998). Adverse event severity was defined as: death; life-threatening; severe; and minor. There were 1015 subject-entries (77% of whom were regular volunteers) but no death or life-threatening event was noted. A total of 43 events were rated as severe, including what the authors describe as *nine worrying cases – six malaises with loss of consciousness, one atrial fibrillation, one hyperthyroidism and one bicytopenia* (0.89%; 9/1015). The "worrying" sub-category was defined by the authors as a "severe" event that was additionally "serious" (as per the standard regulatory definition) or "stressful" (such as loss of consciousness).

Another source of information is data collected by the Association for Independent Clinical Research Contractors (AICRC). The AICRC was a trade association of European Phase 1 Units which was founded in the mid to late 1980s. Its primary purpose was to implement a system of self-regulation by sharing best practice and setting standards, including two-yearly inspections of each unit with certification of satisfactory standard. Part of this Best Practice involved sharing SAE information annually on a confidential basis. Unpublished data recorded during the years 1991 to 1998 inclusive by the AICRC provide a crude measure of SAE risk, and have the advantage of covering all studies in the participating units over this defined timeframe. During this period there were over 65,000 subject-entries into studies conducted at participating Phase I units with the percentage of subject-entries with an SAE ranging from 0.12% to 0.30% per year. Overall the entire eight year period, the percentage was 0.20% (128/65,205) – about 1 in 500 subject-entries. (The AICRC was disbanded in 2003 due to a fall in members and the activities have now been incorporated into the Clinical Contract Research Association (CCRA) – although this does not now include the sharing of data on SAEs. The CCRA's remit is to represent the professional views and interests of its members to government and overseas agencies and it is open to all CROs rather than just Phase I Units.)

In terms of data that should be become publicly available in the future, The Association of Human Pharmacologists in the Pharmaceutical Industry (AHPPI) is currently conducting a survey to identify all SAEs occurring in Healthy Volunteer studies with an Investigational Medicinal Product (IMP) - that is, studies requiring a CTA - conducted within the UK. The survey will collect the absolute numbers of each SAE by subject and by study (including recurrences) and the total number of subjects taking part in the study. The survey will include both biological compounds and small molecules. The reference period for the survey will be 1st January 2004 to 31st December 2005 and will include all eligible studies - even if no SAE occurred - with study eligibility determined by the time of ethics committee submission. There is no requirement for disclosure in relation to mode of action of the IMP such that anonymity can be maintained. Collected data will include all SAEs - whether a drug relationship is concluded or not - and any information on the likelihood of causality. To avoid double-accounting, sponsors will be asked to provide information only on their UK in-house studies and not on studies placed at Phase I units at CROs - the CROs themselves will provide the data from studies conducted at their own units.

While these examples (and associated data) illustrate that the occurrence of serious adverse events in healthy volunteer Phase I studies has been relatively rare (<1% subject-entries), it is clear that the collection, classification and reporting of (and indeed access to) such data is not ideal. The examples illustrate issues with the consistency of event definition and classification such that it is difficult to compare and combine these event data across databases. Information on design (including relative dose) is absent and no statistical adjustment is made for non

independent data generated from individuals who were enrolled in more than one study over a defined period. Furthermore there is no consideration given to basic demographic sub-groupings - such as age and sex – or type of IMP that could affect the risk of harm.

Clearly in order to evaluate ongoing and future risk of harm for healthy volunteers, in the fast moving area of clinical research and development, it is unsatisfactory to rely upon data reported 10 or 20 years ago – or even regular Phase I unit specific summaries. The most obvious solution therefore is to direct attention to the safety data that are routinely reported to regulatory authorities. These data naturally represent the most complete, current safety data available and are an obvious source from which to estimate the risk of harm. In the following section, the current regulatory procedures for SAE reporting are summarised.

### *5.2. Current SAE reporting requirements in the European Union*

Since the implementation of the European Union Clinical Trial Directive in May 2004, all 25 member states - as well as the 3 countries included in the European economic area – must follow the same reporting requirements for SAEs occurring within Clinical Trials carried out for an IMP (Enterprise and Industry Directorate-General, 2006). An SAE is formally defined by the regulators as any untoward medical occurrence that at any dose results in: death or is life-threatening; a requirement for inpatient hospitalization or the prolongation of an existing hospital stay; a persistent or significant disability or incapacity; or a birth defect in a study participant's child (ICH E2A). This definition enables events to be classified in a consistent manner and is the first step in the achievement of standardisation. A further important distinction is between adverse events and adverse drug reactions. ICH E2A defines and adverse event as *any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with treatment*. An adverse drug reaction has the additional qualifier that *a causal relationship between a medicinal product and an adverse event is at least a reasonable possibility* and therefore drug reactions are a subset of events (International Conference on Harmonisation, 1995). (It should be noted that no standard international definition exits for the term “reasonable causal relationship”, although ICH E2A states that the spirit of the term is that its use is supported by evidence.)

The sponsor of a clinical trial from Phase I through to Phase IV, which has at least one investigator site within a European community, must report all SAEs that occur both within the concerned trial and outside the concerned clinical trial. The responsibility for reporting to the competent authorities and member states, the relevant concerned ethics committees and investigators concerned, lies with the sponsor of the trial. Furthermore, a subset of SAEs qualifies for expedited reporting – these are the suspected unexpected serious adverse reactions (SUSARs). In this context “unexpected” refers to an event never previously reported with the IMP and as such serious adverse reactions are by their very nature SUSARs in first-in-man studies. An event is described as “suspected” if there is a reasonable suspected causal relationship. In effect therefore the expedited reporting of SUSARs is a means of informing regulatory authorities of serious unexpected adverse reactions and enabling this new information to be disseminated to stakeholders swiftly. Once such adverse reactions are included in the updated Investigator Brochure, then subsequent identical SAEs are not now unexpected and are therefore subject only to standard, non-expedited, SAE reporting requirements. Therefore expedited reporting provides a mechanism for updating the risk associated with individual subjects enrolled in ongoing studies and enables a decision to be made in relation to study discontinuation. The reporting timelines are 15 days for non-fatal and non-life threatening SUSARs with further relevant information as soon as possible, and 7 days for fatal or life threatening SUSARs with further information within an additional 8 calendar days. The sponsor

must report to the competent authority, and the ethics committee of the concerned member states, all SUSARs associated with a comparator product in the concerned clinical trial - even if this product is authorised. This may include SUSARs associated with placebo if the reaction is thought to be due to an excipient (that is to say some substance used to formulate a treatment or placebo). Member states have the option to legislate that ethics committees in their particular country need receive expedited individual reports only of SUSARs that occurred in subjects who have been recruited in the concerned trial in that member state. In this situation all SUSARs from other member states and third countries are periodically reported, at least every six months, as a line listing.

In addition to the expedited reporting, sponsors are required to submit annually throughout the duration of a clinical trial, or on request, a safety report to the competent authority and the ethics committees of the concerned member states, taking into account all new available safety information received during the reporting period. The annual safety report has to contain an analysis of the subjects' safety in the concerned clinical trial, a line listing of all suspected serious adverse reactions - including all SUSARs - that occurred in the concerned trial whether within the Union or in a third country, and an aggregate summary tabulation of suspected serious adverse reactions that occurred in the concerned trial. The reporting timeframe for annual reports starts with the date of the first authorisation of the concerned clinical trial by a competent authority in any member state The sponsor has to submit annual reports within 60 days of the data lock point. In the case of a first-in-man trial, and other short term metabolism case studies, the safety report has to be notified within 90 days of the end of the trial, together with the notification of the end of the trial. This report has to contain at least an analysis of the subject safety and line listings and, if appropriate, aggregate summary tabulations.

### *5.3. Improving the reliability and availability of data to assess the risk of harm*

As pointed out in Chapter 2, a systematic and prospective method of collection, storage and dissemination of Phase I healthy volunteer safety data – including first-in-man studies - is required. In the current technology driven environment there is little excuse for not having broader access to complete *real time* safety data and although the argument of preserving commercial confidentially should be addressed, this should not be at the expense of providing stakeholder access to important safety information.

In terms of who should hold the data, it is clear that the Regulatory Authorities are best placed to maintain a comprehensive Phase I safety database since, as discussed in Chapter 2, SAEs from all studies (including Phase I) are routinely collected by these authorities. (In the UK, this task is performed by the MHRA – although until May 2004 there was no national database to record the incidence of SAEs and the reporting of SAEs from Phase I studies was not required.) Ideally the data should be combined at a global level, although it could be undertaken at a regional level also. Sibille *et al* have proposed the involvement of World Health Organisation (WHO) to co-ordinate such a task, although the European Medicines Agency (EMEA) or the US's Food and Drug Administration (FDA) might be better positioned for such an undertaking (Sibille, et al., 1998).

The first step to create a Phase I safety database is to specify the data requirements – since, as highlighted in Chapter 2, a useful database requires not only information on SAEs, but also key information on study design, volunteer demographics, etc. The second step is to ensure the systematic collection of these data using standard terminology and definitions. Data should be coded using standard dictionaries and stored using standard database structures with standard naming conventions such that data aggregation and retrieval is straightforward. (Ideally similar

standards could be adopted for animal data to facilitate the comparison of data across species – including humans. This could create a more effective bridge from non-clinical to clinical development.) The application of appropriate statistical methods to account for potential biases in the estimation of event rates – for instance, taking into account multiple subject entry, multiple doses, placebo assignment, study duration etc – is also important in terms of reliable reporting.

The following sub-sections discuss the use of standards in order to create a dynamic database to assess and quantify the risk of future harm in healthy volunteer Phase I studies - including first-in-man.

### *5.4. Key data to collect*

Defining the requirements for data capture will require thorough discussion with contributions from multi-disciplinary teams of experts. The intention here therefore is not to provide a definitive list – rather to illustrate the types of information (in addition to standard SAE information) that would enable targeted questions to be addressed in relation to the quantification of risk for subjects entering future Phase I clinical trials. For instance, questions could include: Is the risk of an SAE occurring higher in female volunteers, or the elderly, compared with other subgroups? Is the risk of an SAE higher in first-in-man studies compared with subsequent Phase I studies? In this respect information on design (a point raised in section 0) is important - as is subject level demographic information. In our view, the data types (or categorisations) listed below represent variables of obvious importance, and would provide a basis for a more thorough and detailed multi-disciplinary discussion:

- Subject identifier
- Age, sex and race of subject
- Healthy volunteer or patient
- First-in-man – or subsequent Phase I study
- Active compound (biologic, chemical, first in class, etc) or placebo
- Dose relative to the starting dose in the study
- Study design, including single or multiple dose, parallel group etc
- On-target or off-target assessment of SAE (type A or type B)

It is important to highlight the requirement to collect data on subjects in whom serious adverse events were not reported. These data are as important as the data recorded on subjects who have an SAE and not only provide the denominator for crude (unconditional) estimates of risk, but are also required to address more targeted conditional questions. Inevitably such data would not be collected in an expedited manner, but would be provided after study completion.

One often over-looked aspect to consider is the possibility of multiple subject entry to Phase I studies where subjects have been enrolled in more than one Phase I study over a defined period of time. Multiple study entries from the same individual are not independent and statistical adjustment is required when aggregating data. Phase I protocols typically specify in the exclusion criteria the minimum allowable washout period from a previous study before a subject is allowed to participate in the given study. The standard applied is often a minimum of three months between trials with an IMP and a maximum participation in three trials during the course of a year. (These criteria are to protect the over-enthusiastic volunteer and are driven mostly from the requirement for blood volumes taken during a Phase I trial which typically amount to between 300 and 500ml.) Many organizations also limit the number of studies in their own unit that a volunteer may participate in during the course of a year, but there is no compulsory central

registry to monitor multiple entries across different Phase I units. There is however a voluntary registry that is compliant with Data Protection Laws called The Over-volunteering Prevention System (TOPS). This was set up in 2003 and it is our understanding that around ten Units in the UK are currently members. The registry works on passport number or National Insurance number, and members enter information on the volunteers and can look up the same information on potential volunteers to check whether they have recently been dosed elsewhere. It is our recommendation that responsibility for the maintenance of this registry is taken on by the MHRA in the UK and that participation in the registry becomes mandatory. It is also our recommendation that it is requirement for Phase I protocols to state explicitly the recommended minimum washout period from previous studies before a volunteer is allowed to participate in the given study – a clinical and scientific rationale being provided in each case.

### *5.5. Standard terminology*

Standard global regulatory definitions currently exist for the definitions of adverse event severity, causality and the classification of "serious" (International Conference on Harmonisation, 1995). In this respect, all data reported to regulatory authorities will use standard terminology and enable data to be compared across IMPs.

### 5.*6. Coding of Terms*

The so-called Medical Dictionary for Regulatory Activities (MedDRA) provides a global regulatory standard for the classification, sharing and reporting of medical information – including adverse events. Developed in the early 1990's as part of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) initiative, MedDRA is now widely accepted - and indeed its terminology was adopted by the FDA in November 1997 for adverse event reporting. It is appropriate for all phases of drug development – including devices. Worldwide adoption of this standard would facilitate the merging of databases originating from different regions and countries.

### *5.7. Development of database standards*

The Clinical Data Interchange Standards Consortium (CDISC) is currently establishing standards for the exchange, storage, archival and submission of clinical data (CDISC Clinical Data Interchange Standards Consortium, 2006). The mission of CDISC is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. Two important standards are the Operational Data Model (ODM), which describes the format in which clinical data are stored, exchanged or archived, and the Submission Data Model (SDM) that describes the format for submitting data electronically to the FDA.

An important advance in the development of standards is the creation of clinical research protocols in machine-readable format. In principle such a development enables the linkage of key protocol data (specifically design aspects) with subsequently reported safety data during both the ongoing reporting of safety data and on trial completion.

Again the worldwide adoption of these standards would facilitate the sharing and merging of data across regions and countries.

### *5.8. Conclusions*

5.8.1. An opportunity exists for the MHRA to take a lead role in developing a Phase I safety database to facilitate the quantification of risk of harm. Indeed in most cases a mechanism already exists for the collection of the required data. It is also our

recommendation that MHRA should take responsibility for maintaining a central registry of participating volunteers in the UK. This would be used by Phase I units to ensure individual volunteers are not over-volunteering into Phase I clinical trials. These data would then be combined with safety, demographic and design data to form a Phase I safety database amenable to reliable statistical analysis.

5.8.2. The database should adopt current database standards in relation to terminology, coding, structure etc, and should be comprehensive enough to enable targeted questions to be addressed. It is our recommendation that that database should be accessible to stakeholders outside of the MHRA and by adopting standards have the potential to be merged with data from other regulatory authorities. The MHRA should undertake to generate regular reports of their own on the database and make these reports publicly available.

# 6. Preclinical studies and in vitro & in vivo issues

In this section we consider how pre-clinical studies should be conducted and used in order to provide useful and safe guidance for proceeding to first-in-man studies.

## *6.1. Introduction*

With the benefit of hindsight, it would appear that those choosing the starting doses in the TGN1412 study, based upon in-vitro and in-vivo pre-clinical data, were overly confident in the quality of the conclusions drawn from the pre-clinical experiments. This is apparent in the Investigator Brochure where, on several occasions, the absence of any meaningful effect is taken to be a green-light. This reflects an underlying statistical method of failing to disprove the null-hypothesis of no-effect. In general, for safety issues, the null hypothesis should be formulated as a real safety effect which it is required to disprove in order to assume safety. This mitigates against the possibility of erroneously drawing a conclusion that a dose is safe to administer, simply because of lack of statistical power. In epidemiology a well-known rule of thumb is that if $n$ subjects have produced 0 cases the upper 95% confidence level for the true rate is approximately $3/n$% (Hanley & Lippman-Hand, 1983).

The main source of information available to us describing the pre-clinical work performed with TGN1412 is the Investigator Brochure. Three aspects of this document cause us concern. Firstly, the investigator brochure briefly describes a series of in-vitro and in-vivo experiments, whose quantitative results are reported in a largely qualitative manner. Secondly, whilst individual studies are reported, there is little or no overview of the results across the programme of pre-clinical work. Thirdly, there is little evidence of statistical science in either the design or the analysis of individual experiments and still less in the design of the programme of experiments. This is not to say that statistics played no role; perhaps it did, but there is little evidence.

This chapter suggests where statistics could make a difference in future programmes. There are probably some easily achievable changes that could be made, whilst, in other aspects, it is likely that new statistical methods may require to be developed.

## *6.2. Purpose*

In order to optimise any system, it is crucial to be agreed upon the objective or purpose of that system. The system of in-vitro and in-vivo work which is conducted pre-clinically is to assure safety. This includes those studies or experiments which, whilst conducted primarily to investigate aspects of efficacy, can foreworn the sponsor as to likely problems of excessive pharmacological effect. This is what was observed in the TGN1412 study.

Preclinical work allows the sponsor to make predictions about the safety of administering one or more dose levels of a compound to humans and for these predictions to be made with acknowledged levels of uncertainty. This purpose not only encompasses the experimental work to be undertaken, upon which much focus has historically been placed, but also actively incorporates the translation or prediction step, where quantification is markedly more complex. It also requires the overt acknowledgement of the precision or lack thereof, with which such predictions can be made. Assessing uncertainty or precision particularly requires statistical thinking and, ideally, the involvement of a professional statistician.

Similar situations pertain in other areas of human safety and we should look to these areas to confirm that the methods used in the development of new medicines are best practice both statistically and ethically. As an example of a statistically more advanced treatment of safety, we

could look to the area of food safety. Here, for example, one finds that estimates of toxic levels of pesticides are obtained from experiments, together with uncertainties inherent in the estimation. The amount of likely exposure of man to these pesticides when pesticide-treated food is consumed, is calculated using conservative figures (e.g. for amount consumed) and with distributions (e.g. weight of consumer; uncertainty in effect given a particular dose). These distributions are then combined to give likely overall distributions of pesticide exposure, from which probabilistic statements about safety relevant to a "typical individual" may be made. Uncertainty about maximum exposure is also considered. This appears to be rather more advanced than methods currently in use in drug development.

## 6.3 Proposed Method

### *6.3.1. Use experimental in vitro/in-vivo systems to explore the science*

Experiments should always be designed with a clear objective, to estimate changes or to disprove the null hypothesis of the existence of a certain level of adverse effects. Studies should be sized, ideally, so that conclusions may be drawn at the per-study level. However, recognising constraints upon animal testing, consideration should also be given to sizing studies with a view to optimising an overall meta-analysis or evidence synthesis exercise.

Alternatively, consideration should be given to using a Systems Biology approach, involving the detailed modelling of kinetics at the cell or receptor level. This is an evolving area and the resulting models tend not to be reported statistically so this approach, as currently adopted, will not readily yield estimates of uncertainty.

The mechanism of action and biological activity of a new biopharmaceutical are examined at this stage of drug development. The results have an impact on all further studies, for example on the initial dose selection for first-in-man (FIM) phase or they could even suggest termination of the study. Among many important properties of new biologic, the receptor affinity, receptor occupancy or cross-reactivity are determined. Understanding the receptor-occupancy-response relationship is important for the dose selection (see for example Early Stage Clinical Trial Taskforce – Joint ABPI/BIA Report, pages 15, 16). Another important issue is the concentration-time relationship as a dose response in vitro and in vivo. Modelling such relationships should include modelling of a random error structure (distribution, variance, possible correlations). Model checking procedures should be an integral part of the modelling; for this to be accomplished replications of the observations are necessary, and must be planned.

### *6.3.2. Quantify outcomes, including confidence in results*

It is crucial always to express results together with their uncertainty. Only then can one start to perceive weakness in inferences and the potential for mis-placed confidence. Point parameter estimation is only partly informative and without the knowledge of the precision of estimation it is, in fact, of little use. Each estimate should always be presented together with its standard error. Also, whenever the underlying distribution of the data is known or known approximately, confidence intervals should be given for all parameters of interest.

Whilst it is greatly preferred that assessments of precision be data-driven, this may not always be feasible. In situations where it is impossible to compute meaningful estimates of precision, the sponsor/trialist/experimenter should be required to express a "post-hoc" view of the precision of the experiment in terms of "whilst the observed value of the parameter is XX, the true value might be as low as LL or as high as UU". This is an equivalent process to eliciting a Bayesian prior from an expert, except applied as a pseudo posterior distribution (O'Hagan, 1998).

### *6.3.3. Extrapolate to humans, building in uncertainty in experimental results and uncertainty in extrapolation*

In the very simplest, best understood paradigms, for example those involving allometric scaling, it is common practice to take “the” animal-based result and multiply by “the” appropriate allometric scaling factor. The uncertainty inherent in the estimation of the animal result, if expressed, is not multiplied up to give uncertainty in the human prediction. Neither is the uncertainty in the allometric scaling factor used to reflect the inherent uncertainty in translating across species. A better procedure would involve both the sources of uncertainty and the estimated effect in animals in producing an estimated effect in humans together with realistic uncertainties. In reality, most translation issues are less well developed and the translation step to human carries much unexpressed uncertainty together with the inherent uncertainties in the animal estimates.

In the ideal situation, the science would be so well progressed that clear relationships between in-vitro and in-vivo experiments and likely outcomes in humans would be well understood and quantified. However, it is acknowledged that this is rarely the case in situations where novel molecules are being developed or where the molecules are so particular to humans that in-vitro and in-vivo preclinical experiments are hard or impossible to design informatively. In such cases, it is likely that any results obtained will be able to yield only an imperfect prediction of effects. Efforts should be made to quantify the uncertainty of such predictions, even be it by reference to experts in the area on account of an absence of solid data.

In designing an experimental system, it is necessary to establish how to translate the results from the system into predictions for humans. This may be based upon biological modelling or statistical modelling or even heuristic models. However, with increased use of such models, the translation method and the precision with which translation is effected should be regularly updated and published. It is clearly not in society's interest to maintain secrecy about the performance of safety-related experiments.

In situations where novel experiments/endpoints are being used, it should be possible to establish from experts at least their extremes of belief in the translatability of the results. This might be established through careful questions such as are used to establish individual priors, when working in the Bayesian framework.

Types of uncertainty/confidence

1) Confidence in Purpose: confidence that the experimental procedure is capable of accurately measuring or assessing the thing it is designed to assess, as expressed typically by sensitivity/specificity or similar measures. (Sensitivity is the probability of labelling a case positive given that it is positive and specificity the probability of labelling it negative given that it is negative.) In a well-established system, confidence can rise to “high” with repeated use and further refinements. In novel or poorly-understood areas, confidence must necessarily be lower.

2) Confidence in Translation: confidence that the results of this experiment translate to something meaningful in man. This must be expressed as low until sufficient evidence has accrued to show good translation. Even then, in the case of novel mechanisms or situations, confidence in translation should again drop.

3) Variability in the results from an individual run of an experiment. Where an individual run produces results exhibiting less precision than usual, this should be propagated into

the confidence about the experimental procedure and the translation step. Whenever possible, the individual runs should be repeated to give a better understanding of the variability.

Following this scheme, each experimental design or suite of experimental designs could be classified according to the its "procedural confidence" (1, above) and its "translational confidence" (2 above). This could be expressed in a table of (Low/ Medium/ High) x (Low / Medium/ High) for the system, as illustrated in Table 6.1.

Entries in this table are examples of how the Confidence in Purpose of an experiment and the Confidence in Translation of the results of that experiment interact to give an overall confidence. The overall confidence is indicated according to the following colour scheme:

| **Overall Confidence** | **Low** | **Medium** | **High** |
|---|---|---|---|

The overall confidence should be expressed as the "less-confident" of the two assessments for Confidence in Purpose and Confidence in Translation. Thus, a well established experiment (High Confidence in Purpose), with little Confidence in Translation (Low) would be regarded as Low overall confidence. An overall classification would be reduced if the variability observed were distinctly greater than usual (precision decreased). In fact, variability of the results (categorised as low, medium, high) could be added as a third dimension of the table. Then the overall "high confidence" would only be possible to state if the two main factors were at the "high" level and the variability at a "low" level.

Alternatively put, it is only possible to achieve high overall confidence if:

1. we have high confidence in the experimental procedure itself
2. we have high confidence in the translatability of the results of the experimental procedure
3. the random variation observed in the particular experiment conducted was consistent with that observed historically

### *6.3.4 Choose the starting dose*

The starting dose should be chosen with regard to the uncertainty both in the outcomes of the various in-vitro and in-vivo experiments and, most importantly, with regard to the uncertainties in extrapolating these results to human. The doses chosen should have a suitably low level of risk attached.

Various parameters, denoted here by a vector $\boldsymbol{\Psi} = \left(\Psi_1, \cdots, \Psi_p\right)$, such as pharmacokinetic (PK) parameters, affinity, receptor occupancy, relative potency in animal versus humans (c.f. Early Stage Clinical Trial Taskforce – Joint ABPI/BIA Report, page 20) and determined during the pre-clinical studies are taken into account for assessing the starting dose. All of them, combined in a function, give an indication for the initial dose, which can be written as

$$Dose\,1 = f\left(\boldsymbol{\Psi}\right).$$

Were the values of vector $\boldsymbol{\Psi}$ known exactly, it would be only an issue of establishing the form of the function $f$. Here however, we only have estimates of the parameters, so we can only get an estimate of the initial dose. Its precision will obviously depend on the precision of each $\Psi_i$. Examination of a distribution of the dose estimator by an analytical way (if possible) or numerically (using methods like the bootstrap) would indicate a range of plausible initial doses.

**Table 6.1: Illustration of the use of Confidence in Purpose and Confidence in Translation**

| | **Confidence in Translation** | | |
|---|---|---|---|
| **Confidence in Purpose** | Low | Medium | High |
| Low | **Novel experiment and no experience of translation of endpoint** | | **Novel experiment, but good experience with translating this endpoint (when obtained from alternative experiment)** |
| Medium | | **Experiments becoming established, with some occasional refinement. Some, limited, experience of successful translation of endpoint** | |
| High | **Established experiment, but little or no experience with translating results to man** | | **Established, well-understood experiment or system with well understood translatability** |

Which starting dose to choose for the first-in-man (FIM) study would then be decided based on the distribution, the overall level of confidence (as described in 3.3) and on other aspects, such as for example, batch variability.

A well-established, well-understood suite of preclinical in-vivo/in-vitro experiments, of relevance to a current drug candidate, would have associated with it significant amounts of information regarding safety in humans. The starting dose for such a current drug candidate could be chosen with some confidence, based upon the quantified responses in the pre-clinical experiments and the remaining uncertainty in translation.

However, a less well-established or less-well understood suite of experiments would have significantly greater uncertainty associated with them, usually resulting in a lower starting dose being adopted for the same level of risk. Special caution is required with biologics, as these are less predictable - a small dose may have a strong effect if, for example, there is high affinity for the target (see Early Stage Clinical Trial Taskforce – Joint ABPI/BIA Report, page 15).

*6.3.5. Understand the whole process, then optimise*

Once we understand that we are dealing with a suite of in-vivo and in-vitro preclinical experiments and that we wish to combine the information from them in predictions of effects on humans, we can start to optimise the design of the suite of experiments and the design of the

experiments themselves. While we consider each experiment in isolation, we miss the opportunity of formally combining information to yield better quality decisions.

#### *6.3.6. Insufficient Methods*

It is insufficient to ignore the variability in results from experiments.

It is insufficient to assert that observed effects in one species are of no concern without the experimental data to support the translation of the effect to human and without knowledge of the consequences of such an effect in humans.

It is insufficient to work with a paradigm based upon, "absence of any observed effect means that it is safe", especially when working with extremely small numbers.

### *6.4. Statistics*

Where can statistics help, explicitly?

#### *6.4.1. Modelling and simulation*

There is impetus towards increased use of mechanistic and stochastic modelling in both the pre-clinical and clinical portions of drug development (Stanski, Rowland, & Sheiner, 2005). Modelling should always be viewed together with design and analysis as they are all dependent on each other. Modelling the observations will depend on the plan of experiment, on randomisation of the experimental units, knowledge of possible stochastic effects (such as random errors, “population” variability), but also on the knowledge of the underlying biological/chemical process. Where there is no reliable knowledge about the kind of model for the process then the model is derived empirically, that is, it is determined based on the observed data only. The design of experiment needs to allow for gathering all the necessary information for the model based analysis to be possible.

Mechanistic models (for example, solution to sets of differential equations or Physiologically-Based Pharmaco-Kinetic) allow not only for more informed design and analysis but also for simulation of otherwise time- and cost- expensive experiments. Such efforts are already undertaken within the pharmaceutical industry, in part due to the reinvigoration of the 3 R’s initiative (“Replacement, Reduction, Refinement”) which espouses basic principles guiding animal use in research, teaching and testing.

The advantage of building a mechanistic model is that it can be used to try out alternative scenarios with little, or no, further experimental work. For example, the uncertainties in parameter estimation described previously can be explored through the simulation of possible outcomes. Even without stochastic elements, sensitivity analyses may be performed by varying parameter estimates over a credible range and recording the changes observed in the endpoint of interest. This allows the modeller/experimentalist to “observe” the extremes of the possible outcomes based upon their experimental work and should prove enlightening in most circumstances. Where some information on random errors is available and they may reasonably be assigned an approximate distribution, stochastic simulations may be conducted to give a distribution of likely outcomes, allowing probabilistic statements to be made concerning their likelihood. Resampling methods can be used when no information on the underlying distribution is available.

Without access to the experimental data, we conjecture that this modelling and simulation approach, applied to the pre-clinical data, would have given much greater insight into the range

of possible outcomes to be expected in the TGN1412 clinical trial, before the protocol was finalised. At worst, the trialists would have been better informed.

This approach is being championed in Europe by “BioSim” – an EU sponsored Network of Excellence in Biosimulation [www.biosim-network.net].

#### *6.4.2. Evidence Synthesis*

Taking the preclinical programme, or suite of experiments as a whole, we advocate the use of statistical methods from evidence synthesis and meta-analysis to help formalise the accumulated and accumulating information. This requires the provision of suitable estimates and uncertainties from all experiments (in-vitro and in-vivo) on the original (experiment) scale of measurement. Importantly, this also requires some translation of each estimate and uncertainty to the human domain, at which level meta-analysis techniques may be of some use in combining results. It is fundamentally important to incorporate the uncertainty of the translation step for individual experiments into the uncertainty of the estimates in the human domain.

#### *6.4.3. Bayesian approach*

Many of the results from preclinical in-vivo and in-vitro experiments are conditional upon the exact model being used and the exact endpoint being studied. Additionally, we desire to make predictions in the human domain in the form of probabilistic statements. Taken together, these factors strongly suggest the use of the Bayesian framework for this area. This also allows the incorporation of vague information from other sources (literature, conference reports) and from professional opinion (individual prior elicitation).

#### *6.4.4. Links to Surrogate Endpoints*

Assessing the quality of translation is very close to the problem of "validating" surrogate endpoints, where a novel surrogate endpoint is "designed" to predict the effect of a "true" clinical endpoint. It would be of value to examine how the methodologies established by Prentice (Prentice, 1989) and developed by others (Buyse & Molenberghs, 1998)might be adapted or extended into the preclinical safety biomarker setting.

#### *6.4.5. Optimise the programme of preclinical studies with respect to the quality of predictions*

If the primary purpose of a programme of in-vitro and in-vivo experiments is as previously defined and if the choice of dose is to be made in a defined manner following a systematic review of the accumulated information, it should be possible to work back to optimise the design of the programme and the design of the individual experiments in order to minimise animal use, risk to human and cost.

#### *6.4.6. Statistical reporting*

It should be standard practice to always give full statistical evidence supporting any conclusions drawn based on any statistical analysis. This should include:

- Description of the design (also including number of observations, missing values etc).
- The model and all the assumptions made on the prior parameter distributions, random error structure, etc.
- Kind of analysis performed (estimation, testing together with the significance levels).

Also, it is important to be careful with wording the conclusions, as many of the statistically meaningful words are often used in colloquial language and are easily misinterpreted.

It is insufficient to present only point estimates of parameters of interest without any information on the way the data were obtained, how many data were used and the variability of the data.

Several places in the TGN1412 Investigator Brochure give point estimates with no indication of variability. We found the reporting there to be deficient in many of the above respects.

### *6.5. Conclusions*

We propose that the purpose of the body of pre-clinical work is to allow the sponsor to make predictions about the safety of administering one or more dose levels of a compound to humans and for these predictions to be made with acknowledged levels of uncertainty. We have indicated that existing and developing statistical methods, if applied to such a body of work, can give much greater information on the credibility of such predictions. We conjecture that, had such statistical methods been rigorously applied in the TGN1412 case, the confidence in the starting dose (if any) would have been less and different outcomes might have resulted.

# 7. Design and conduct of First-in-Man studies

In this section we discuss the key issue of appropriate design of first-in-man studies. Such studies have a number of aims and, as we point out, this may involve some compromise in designs. However, whatever aims are given greatest emphasis, the link between design and intended analysis is key.

## *7.1. Terminology*

When biologists or medical people need to describe a new concept, they invent a new word. In contrast, statisticians and mathematicians take an existing word and give it a special meaning. In this chapter, such words are shown in bold face at their first occurrence.

Experiments on human subjects are usually careful about a range of issues, including the following.

**Randomisation** This means that treatments are randomly allocated to people, subject to the overall proportions on each treatment being correct. This should avoid any treatment being given favourable status. Likewise, the laboratory technicians analyse samples in random order, so that they do not analyse all of one sort at the end of the day, when their attention might be lower.

**Blinding** This means that the human subject does not know whether they are getting placebo or not, nor which dose of the drug they are getting, so that their reactions are not purely psychosomatic. In **double-blind** trials, the investigators are also ignorant, so that they are not subconsciously biased in their assessment of the people. Likewise, the laboratory technicians do not know which dose they are analysing, so they have no subconscious wish to make certain samples have similar results.

**Blocking** The effects which we measure may be affected by extraneous factors in which we have no direct interest. For example, if two trial centres are involved, all the human subjects at one may react better than those at the other. This is called a **centre effect**. Similarly, if some subjects are treated in winter and the others in summer, or if the same subjects change drugs between the seasons, then there may be a **period effect**. A common strategy in experimental design is to allocate treatments in such a way as that all treatments are applied to the same number of units in every block. For this reason, medical statisticians have come use the word **block** to mean a group of patients for which the allocation is constrained in this way.
There may be other factors, such as sex, over which we have no control, but whose effects are of interest.

In the context of clinical trials, some types of block, such as centre and sex, are called **strata**. A **cohort** is another type of block: a group of people who are recruited into the trial, and treated, at (roughly) the same time.

Where possible, the study is designed in the knowledge that there are blocks (of whatever type). For example, the pattern of treatment allocations may be made deliberately the same in each block. Then the data are analysed in a way that incorporates the knowledge of which subject belongs to which block.

### *7.2. Statistical input into the protocol for first-in-man trials*

Phase I and early phase II studies are generally thought of as exploratory or "learning" whilst late phase II and phase III are thought of as "confirming". Generally in phase I and early phase II, one expects to be exploring the data in order to learn. One then uses the intelligence gained to help design good phase III studies, where learning is "disallowed". We are now confirming afresh, for regulators, what we believe we already know. So, it is quite acceptable, indeed necessary, to have some flexibility of approach in this early stage. However, we should indicate our intentions in the protocol!

There can sometimes be confusion, as the term "Phase I" is sometimes taken to be synonymous with "Healthy Volunteer, pharmacokinetic studies", some of which may be among the last studies to complete before a new medicine is submitted to regulatory authorities. In general, most pharmacokinetic statistical analyses have been standardised for some time; for example, no-one would think twice, now, about the need to analyse log transformed Area Under the Curve (*AUC*) and Maximum Observed Concentration ($C_{max}$). Similarly, methods to determine bioequivalence are now nearly non-negotiable and are also widely accepted as very sensible. This results partly through regulatory guidance and partly through accumulated experience of those working in the field.

The statistical methods section of a typical protocol with a pharmacokinetic component would include something along the lines of "log transformed AUC will be subject to an analysis of variance (ANOVA) suitable for the 4-period crossover design. Estimated means, differences in means and 90% confidence intervals will be presented on the log scale. Estimated geometric means, ratios of geometric means and (back-transformed) 90% confidence intervals will be presented on the nominal scale." These statements would be backed up with further detail (such as the exact model, period or carryover terms, details of graphical presentations and tabulations etc) in the Statistical Analysis Plan. This document must be signed off internally before the database is frozen. After this point, no data may be modified or added. Only then may **unblinding** take place: that is, the investigator can see which actual dose or drug was administered to each subject. The various different approaches to pharmaocokinetic data are catalogued in company-specific guidance to ensure consistency and to save "re-invention of the wheel".

When dealing with data other than pharmacokinetic data, fewer standards exist. In cases where we have rather little experience with the endpoint in a particular population or with the endpoint per se, we would include statements such as "The endpoint will be subject to a multiple regression to establish the influence of age, sex, treatment and disease severity. Dependent upon the results obtained, further exploratory analyses may be conducted to understand the data more fully. Should the assumptions of the method not be adequately satisfied, alternative procedures may be used and will be fully documented." A detailed Statistical Analysis Plan may not be attached to the protocol in such situations, because it may not be ready until the trial is well under way, but it will be appended to the protocol before unblinding. However, in view of the relative rapidity of phase I studies we could consider making it mandatory that a full Statistical Analysis Plan be available as part of the protocol.

The balance is between the amount of detail to go in the protocol and the amount of detail to reside in the Statistical Analysis Plan. To put every detail into the protocol would be considered inappropriate and unnecessary by many, including some members of ethics committees.

The experience of the independent statistical consulting companies is that the large and established pharmaceutical companies provide a very detailed methodology section in protocols - including their phase I protocols. Smaller companies are less likely to do so. The latter may have no in-house statistician (or only a junior one) and may not seek external advice prior to getting their protocol approved. Of course, many small companies do seek advice (or have an experienced internal statistician) and their protocols are indistinguishable from those of the large companies.

There might be a wider issue that a disproportionate number of phase I studies are sponsored by small/new/virtual companies. Their aim is most likely to show proof of concept/principle and then sell the drug off to large pharmaceutical company, who will then take it into late phase II and phase III. One could speculate therefore that in practice a lower proportion of phase I protocols receive due statistical attention (compared with phase III).

## *7.3 The TeGenero first-in-man trial*

### *7.3.1. Design employed*

The protocol states that "the primary objective of this first-in-man trial is to establish the safety and tolerability of TGN1412 in man" (p. 16). However, as will be discussed below, the trial was not particularly well-suited to investigating either safety or tolerability, nor even for estimating pharmacokinetic parameters, such as how long TGN1412 stays in the body, its maximum concentration or the time after dosing that this is achieved.

Several pharmacokinetic parameters were to be derived from serum concentration of TGN1412 (Parexel, 2005)(p. 49, sec. 9.2). It is said in this section that "All TGN1412 profiles will be presented graphically as individual curves." One would expect to get some indication here of how these curves would be obtained. Would they be based on a statistical model of concentration, such as for example a one-compartment model, or would they be obtained by some other method? Also, there is no mention of random errors of observations. These curves are important for precise evaluation of the PK parameters. Calculating their values based on a very rough profile, such as a linear point-wise profile, may be very inaccurate if the sampling times are widely spread. It might give a reasonable parameter evaluation if the sampling followed a dense time grid. Further in this section we read "The maximum serum concentration ($C_{max}$) and the corresponding time ($t_{max}$) will be read directly from the serum concentration data." This makes the sampling times even more important. On the page 31 of the Protocol we find the planned Day 1, Post-Dose procedures, where the last bullet point gives the blood sampling times for PK as 1, 2, 4, 8 and 12 hours after start of infusion. These times may well miss the peak concentration. Sampling times might have been better chosen if appropriate pre-clinical information had been used.

Another important issue is how to interpret the separate, probably quite different, profiles. The statistical model approach would give the possibility of treating some of the PK parameters as random variables and so of estimating their mean values and their variances. This would give a more comprehensive interpretation of the expected values of the parameters as well as of the variability among the subjects. Given some prior information of the PK parameters, it would be sensible to derive an optimum sampling plan with respect to parameter estimation.

Safety studies were also poorly designed. For example, it was known from the animal studies (Investigator's Brochure, p. 46) that after administration of the first dose of TGN1412 in cynomolgus monkeys, peak serum concentrations of cytokines IL2 and IL6 (inflammatory mediator) were observed as soon as after 2 hours. However, the anti-inflammatory IL5 did not

reach peak until 24 hours after injection. Bhogal and Combes (ATLA 34, 225 – 239, 2006) suggest that this may imply that the initial effect of TGN1412 might have been predicted to be inflammation. TeGenero's plan for blood sampling for PD cytokines was 1 and 4 hours post-dose! Moreover, the samples were to be analysed in a private off-site laboratory, with the effect that PD profiles were not immediately available to guide the clinical management of those volunteers who experienced CRS.

Similarly, the plan for sparse measuring of temperature, ECG and observation of other vital signs might be questioned (Parexel, 2005)(p. 31, sec. 6.2.2).

The design envisaged in the protocol was a **dose-escalation study**. In such a study, doses are introduced in increasing order. We shall call the doses **Step 1**, **Step 2**, and so on, irrespective of the actual quantitative value of the doses. The design was for 32 subjects, divided into four **cohorts** of eight subjects each. The plan was to treat all subjects in a cohort on the same day and then observe them for 14 days before treating the next cohort. Within each cohort, six subjects were to receive the drug at one step higher than the previous cohort, while two received placebo. The planned numbers are shown in Table 7.1.

**Table 7.1 Design given in the protocol**

| | TGN1412 | | Placebo |
|---|---|---|---|
| Cohort | Dose [mg/kg bodyweight] | Number of Subjects | Number of Subjects |
| 1 | 0.1 | 6 | 2 |
| 2 | 0.5 | 6 | 2 |
| 3 | 2.0 | 6 | 2 |
| 4 | 5.0 | 6 | 2 |

However, as said in the Protocol (Parexel, 2005), (p. 49), “The total sample size of 32 subjects is not based on a formal statistical assessment.” One may ask: “Why not?” The next sentence in the Report says that, “… this number of subjects is considered sufficient to achieve the objectives of the study.” Again, one might ask, “how does one know?” Also, the reasons for division of the 32 subjects into cohorts of 8 and then further division into 6 for the active drug and 2 for placebo are not justified.

Randomisation was carried out in groups of four. That is to say, for the first four subjects recruited in each cohort, one was chosen at random to be given the placebo, while the other three received TGN1412. This procedure would be repeated for the second group of four.

Progression from one step to another required observation for at least 14 days for subjects on the current step and approval by the data safety monitoring board of the study. In the event, due to rapid development of serious adverse effects in the six subjects allocated to TGN1412 in step 1, the trial never proceeded beyond this point.

### *7.3.2. Analysis envisaged*

The planned statistical analysis was not described in the sort of detail that would permit clear identification of the method envisaged (Parexel, 2005), (p. 4 and pp. 49 – 51). Reference was made to a statistical package (SAS®) and to a very general technique (analysis of covariance) that can take many different forms. The biochemical analogy would be to refer to a commercial supplier of chemical reagents and then say that an assay would be performed without giving any

details of the assay planned. Failure to mention the technique used for analysis is important because it impacts on choice of design and vice-versa (Bailey, 1981), although it must be admitted that the events that occurred were so dramatic as to render any plans there were for this trial irrelevant. The plan for PK sampling and analysis was not justified and the parameter evaluation not clear. However, this is a common feature of many phase I trials in healthy volunteers.

Apart from allowing for a safety and tolerability assessment for each dose before proceeding to the next step, another advantage of this approach is that the concentration-time curve for each subject can be estimated before the next cohort is dosed. The shape of this curve, including the values of (i) the time to maximum concentration and (ii) the time till the drug is cleared from the body, can inform the choice of time-points at which to measure concentration in later cohorts. However, neither the lab technician analysing the blood samples nor the statistician doing this estimation is blinded as to dose. Of course, there are no PK data for placebo subjects.

TeGenero's Clinical Trial Protocol (page 49) explicitly says that “data of subjects having received placebo will be pooled in one group for analyses”. This masks some important assumptions, as we discuss in Chapter 4.

### *7.3.3. Design for safety*

there are several implications of having employed the design in Table 7.1. The first, to be discussed in more detail below, is that it is clear that it can be suitable in general terms only if the object of the trial is not to confirm safety of the product but to establish tolerability (in fact probably the maximum tolerated dose) and indeed only if it is believed highly probable that the trial will proceed to completion (step 4 in this case).

*The Textbook of Pharmaceutical Medicine, 5th edition*, (Posner, 2005) page 159, states that the objectives of a first trial in humans of a new molecular entity should be to investigate, over a range of doses, (a) safety (does the drug cause harm?) (b) tolerability (does it cause inconvenience?) (c) pharmacokinetics (what does the body do to the drug?) and (d) pharmacodynamics (what does the drug do to the body?) (Posner, 2005). However, such a trial would not usually be envisaged unless the investigators were fairly certain about safety. They reach such certainty because the drug is similar to one in current use, because the drug targets a bodily function which is also targeted by known safe drugs, or because studies in animal species which are known to react like humans give confidence in the safety.

One lesson to be drawn from the TeGenero trial (and given by various other reports, c.f. ABPI Report) seems to be that if

- the drug is a novel agent, or
- it targets a novel aspect of the human body, or
- there are no adequate animal models,

then the first trial in humans must be aimed primarily at establishing safety. Of course, measurements can be made to find out about the other three objectives, but the over-riding concern in the design of the study must be the detection of unsafe compounds and the minimization of risk to human subjects. All, even small, indications of possible adverse reactions to the new drug observed in pre-clinical studies should be taken into account in designing the first-in-man trial.

Although the primary objective is safety, PK is vitally important to understanding this, because adverse events are typically driven by the drug levels. For instance, one might find that one subject who had a serious adverse reaction also had high concentrations of the drug. Then we have to try to understand why that happened. Thus PK measurements are an important part of First-In-Man studies.

For the purpose of assessing safety, it was clearly not necessary to treat six subjects simultaneously with the active treatment. For example, suppose one takes it as a given that six subjects must be studied at a given dose step before proceeding to the next. A more cautious design would have been to randomise the subjects in a single cohort of 12 with six given the active treatment and six given placebo, at intervals to be based on predicted pharmacokinetics and dynamics of the treatment and chosen in such a way that a sufficient time is given for problems to emerge before the next subject is treated. We shall refer to such an interval as a **proper interval**. Also, the blood sampling times and observation of vital signs should be chosen with care, based on the existing knowledge of the drug.

After all, in phase II and III studies, it is usually accepted that patients will (usually) not be recruited simultaneously. There seems to be no compelling reason, apart from speed, cost, convenience, and reduction in variability of PK data, for subjects to be treated simultaneously in phase I studies.

Of course, such an approach would lead to an increase in the number of subjects per dose step from 8 to 12 compared to the design used in the trial of TGN1412. One could envisage unequal allocations such as six versus two. Since, as discussed in 7.3.5, inferences regarding occurrences of (usually) rare, severe and acute reactions do not require concurrent controls, the main contribution of equal allocation is to sharpen inferences about tolerability and, as discussed in Section 7.4.3.2, for this purpose, four versus four is superior to six versus two.

An alternative design would randomise two subjects simultaneously, one to placebo and one to active treatment, but proceed to the next pair only after a proper interval had been observed. If, again, six subjects per dose step is taken as a given, then six such pairs would be needed for each step.

The main disadvantage of a single cohort of 12 (six of whom are randomised to the drug) is that 11 proper intervals are required to complete the dose step because such an allocation makes it possible for consecutive pairs of treated subjects both to be receiving active treatment. In contrast, only five proper intervals are needed for six cohorts of two since each subsequent subject to be given active treatment occurs in a subsequent cohort.

The advantage of the approach using a single cohort of 12 is that it achieves the maximum degree of unpredictability as regards sequences employed, there being $12!/(6!6!) = 924$ possible sequences. On the other hand, six cohorts of two subjects can be arranged in only $2^6 = 64$ ways. However, as previously discussed, the main contribution of blinding in phase I studies is to sharpen inferences about tolerability rather than safety and it will depend on the relative importance of these objectives to what extent one regards larger cohort sizes as being important.

### *7.3.4. Informal analysis*

It may be argued that discussions such as those given above are irrelevant since it is rarely the case that formal statistical methods are appropriate for analysing such studies and human judgement is used instead. The role of human judgement is, indeed, important, as will be

discussed in Section 7.3.5 below when considering the results from step 1 of the actual trial, but is a false opposition to place this against statistical analysis.

Statistics is not *a* theory of inference; it is *the* theory of inference (or a collection of theories) and whereas conventional statistical methods may be too rigid to take account of all relevant information this does not mean that more flexible approaches are impossible. There is also the problem that informal analysis may simply overlook relevant issues. For example, the same sorts of bias in employing veiled designs referred to in Section 7.4.3.2 can also affect informal judgement as to the relevance of safety data and not just statistics based on these data.

### *7.3.5 Implications of the results of step 1 of the trial of TGN1412*

A simple summary of the results of step 1 is given in Table 7.2.

**Table 7.2**. **Outcomes for cohort 1 in the TeGenero trial**

| | | Adverse reaction | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Treatment | TGN1412 | 6 | 0 | 6 |
| | Placebo | 0 | 2 | 2 |
| | Total | 6 | 2 | 8 |

To analyse such a contingency table the conventional mechanism of hypothesis and significance tests is often used. It is assumed that there is no relationship between treatment and outcome and this becomes a **null-hypothesis**. The probability of seeing a result as extreme or more extreme (in the sense of suggesting a connection between treatment and outcome) is then calculated. This probability is known as a **P-value**. There are several possible approaches to calculating such a P-value. Fisher's exact test is one of the most common. If this is used, a one-sided P-value of 0.0357 is obtained. This test is not uncontroversial, because it conditions on both margins, but whereas it was known in advance that the treatments would split six to two it was not known that the side effects would also split six to two. Fisher's exact test assesses the unusualness of the pattern (all six side effects under TGN1412 and none under placebo) given that the six to two split has occurred, but the split itself is suggestive. A test that does not condition on both margins is Barnard's test, which yields a one-sided P-value of 0.0111.

It is clear, however, that neither of these analyses does justice to the result. All commentators are convinced - with good reason - that the drug is highly toxic in the dose given; these fairly modest P-values do not begin to do justice to that degree of conviction (Senn, 2006). This raises the issue as to how such judgements are made.

There are two plausible sources of further information not conveyed by Table 7.2. First, the background knowledge that adverse reactions of the sort that occurred are almost impossible in subjects given placebo; secondly, the timing of the reactions. (The use of timing as a means to judge causality in epidemiology is receiving increasing attention: (Farrington & Whitaker, 2006).)

This, of course, immediately raises the issue, previously discussed in Section 7.3.3, that designs of the sort employed are not particularly appropriate for basic investigations of safety. On the contrary, they are appropriate for investigating tolerability, that is to say the ability of the treatment to produce the sort of dose-limiting side effect such as headache, nausea, sore-throat,

tiredness and diarrhoea that cannot be excluded as possible symptoms under placebo and whose frequency may vary considerably from circumstance to circumstance.

This is not to say that placebos should be excluded from first-in-man studies. It recognises, rather, that the importance of placebos varies according to the purposes of such studies, which may be several.

## 7.4. Design of dose-escalation studies

### *7.4.1. Preliminary issues*

Pre-clinical studies need to establish several important points before any first-in-man experiments are run. The issues include

- Choice of subjects: healthy volunteers or patients. The altered immune system in patients may react to biologics in a different way from the immune system of healthy volunteers. Differences in receptor density between healthy people and patients can cause differences in PK profiles (clearance, half-life); there may also be differences between the two groups of people in the tolerability of the new therapeutic protein. Volunteers aged 18 – 40 are particularly prone to T-cell stimulation (See chapter 3.)

- Initial Dose ($D_1$): The dose level to be administered to the first subject (or the first group of subjects). This is a crucial issue in all drug development studies, but in the case of biologics, it requires more detailed investigation as it may be that a lower dose could have a divergent effect (even enhanced potency) compared with higher doses.
- Method of administration of the drug.

All these issues will have an effect on the first-in-man study design.

If the drug is a novel agent and there is no information from previous experiments in humans, the pre-clinical studies are of paramount importance and should be properly designed and analysed to give as accurate information as possible on the agent's mechanism, toxicity, possible adverse reaction and other issues. (Stanski, et al., 2005) in their report on getting the dose right strongly recommend “model based methods” of study design and analysis. They also advocate adaptive clinical trial design and numerical simulations in early drug development stages.

Based on the pre-clinical information and on the purpose of the further studies, an experiment must be designed with caution to allow for the unexpected.

### *7.4.2. Sequential choice of dose*

Two of the main objectives of Phase I studies are to establish PK parameters and to find a safe (but potentially useful) dose to be further examined for its effect in later phases of the drug development. Below we point out and discuss some important elements of such a trial, though we are aware that the discussion is not exhaustive and we may have missed some other crucial issues.

**Dose 1, Subject 1**

The initial dose needs to be evaluated based on the pre-clinical studies and, in the case of biologics, it is recommended that a Minimum Anticipated Biological Effects Level (MABEL) is established, as well as a toxic dose and a No Observed Adverse Effect Level (NOAEL).
The first subject in the trial is at a higher risk and the administration of the drug needs to be slow and to be carefully observed. In the document *Early Stage Clinical Trial Taskforce - Joint*

*ABPI/BIA Report* it is stated that an intravenous drug administration should last at least 60 minutes and be observed for local reaction at the site of injection. The whole trial may have to be stopped at this early stage if the local reaction causes serious concern.

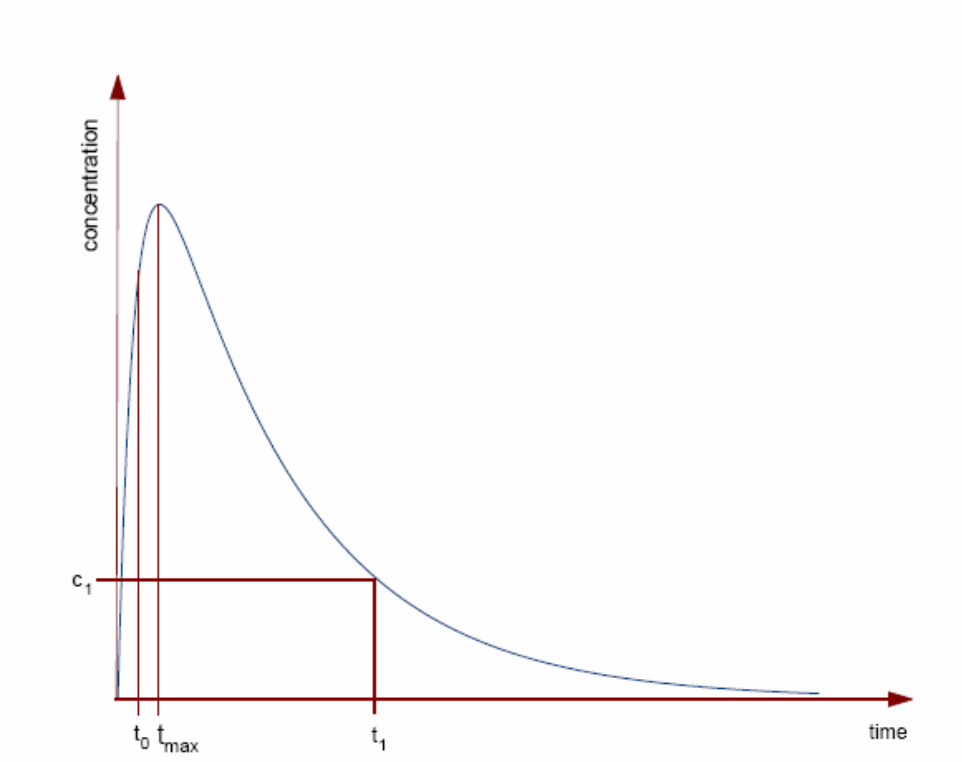



**Figure 1. Concentration versus time: $t_0$ - time of administering the drug, $t_{max}$ - time of maximum concentration, $t_1$ - time when the concentration after the peak has returned to a low level $c_1$, with no adverse reaction being observed. Note that this is not necessarily the picture one would expect for an infusion over a period but is meant to illustrate general pharmacokinetic features.**

An accurate evaluation of the model of drug concentration in blood (plasma), and so of the PK parameters, is important for making decisions such as when to treat the next subject and at what dose. If there is no indication on the kind of PK model, a dense sampling scheme may be necessary for establishing the time-concentration relationship. However, if it is known that, for example, it is a two-compartment model, then a robust optimum design for finding sampling times should be recommended. The theory for optimum design for non-linear models is well developed in statistical literature, see for example (Atkinson & Donev, 1992).

The trial may have to be stopped at this stage, if serious adverse effects are observed.

**Dose 1, more subjects**

Having the observations and the estimates of PK parameters, such as absorption and elimination rates, volume of distribution and also $t_{max}$, $C_{max}$, *AUC*, a decision needs to be made on the next step of the study. The Joint ABPI/BIA Report recommends at least 3 or 4 subjects to receive a given dose level prior to any subject receiving a higher dose. The PK profile of the first subject may help to decide when, at the earliest, to treat the next subject. It is recommended that the interval before the next subject should be at least as long as the anticipated time of peak plasma concentration $t_{max}$ or maximal PD effects. Perhaps, at the beginning of the trial, it would be safer to start treating the next subject at time $t_1$, which is not less than the time needed to attain some

low (safe) concentration after the peak, given that no adverse effects were observed, as shown in Figure 1.

Similar precautions should be taken for the later subjects as for the first subject, particularly because there may be a large variability between subjects in their reactions to the treatment, a variability which at this stage would not be known. Some information about the variability may be obtained from the pre-clinical studies, and it could be used in phase I planning if there are grounds for the information to be reliable (like a good animal model).

**Dose *i*, Subject *j***

The ABPI/BIA Report recommends that a complete PK analysis should be done before increasing the dose and that the time between cohorts of different doses should be at least 24 hours. What "a complete PK analysis" means is not said, neither it is explained why 24 hours.

This still leaves open the question of whether the data should be analysed after one person or after a whole cohort. Likewise, should the next dose be chosen after each subject or after each cohort?

When the treatment is to stimulate one's own immune system the dose-response (toxicity/efficacy) may not follow standard models. There may be unimodal dose-efficacy relationship when low dose levels do not result in efficacy, large doses give toxicity, but there is an optimal dose in-between, which gives the required effect. Such a dose is called the Biologically Optimum Dose (BOD). It may also be that the lower the dose the better or the other way round.

The dose escalation study may be based on the following trinomial outcome
$y = (y_0, y_1, y_2)$,
where $y_0$ denotes the number of subjects where the dose has "no efficacy and no toxicity", that is "no response", $y_1$ denotes the number of subjects where the dose has "efficacy and no toxicity", that is "success" and $y_2$ denotes the number of subjects where the dose has "toxicity", that means "failure" - efficacy or not. When $n$ subjects are treated at dose $x$ then

$$y_0 + y_1 + y_2 = n.$$

The probabilities of each of the outcomes, treated as functions of dose $x$ and of some unknown parameters $\vartheta$, $p_j(x, \vartheta)$, $j$=0,1,2, satisfy

$$p_0(x, \vartheta) + p_1(x, \vartheta) + p_2(x, \vartheta) = 1$$

and are chosen to follow the expected shapes of the responses, that is,

- $p_0(x, \vartheta)$ decreases with dose,
- $p_1(x, \vartheta)$ decreases with dose, or increases with dose, or is unimodal,
- $p_2(x, \vartheta)$ increases with dose.

Such scenarios of dose-response are considered in the recent paper by(Zhang, Sargent, & Mandrekar, 2006). See also (Bretz, Pinheiro, & Branson, 2005). In fact, it might be reasonable to widen the choice of scenarios to allow $p_0$ and $p_2$ to follow other shapes as well.

Various kinds of dose-response model may be used, such as Proportional Odds or Continuation Ratio, to represent the behaviour of these probabilities over a range of doses. Also, various

criteria can be formulated to optimise the dose-escalation design. (Zhang, et al., 2006)suggest two decision functions

$$\delta_1(x,\vartheta) = \begin{cases} 1 & \text{if} \quad p_2(x,\vartheta) < \pi_0 \\ 0 & \text{otherwise} \end{cases}$$

$$\delta_2(x,\vartheta) = p_1(x,\vartheta) - \lambda p_2(x,\vartheta),$$

where $\pi_0$ is a maximum probability of toxicity allowed, specified *a priori*, and $\lambda$ has a value between 0 and 1. The toxicity function $\delta_1$ is used to provide a safe dose domain *X*, that is such a dose range that the toxicity has a chance to occur smaller than the probability $\pi_0$. The function $\delta_2$ is maximised over *X* and the BOD is defined as

$$x^* = \arg\max_{x \in X} \delta_2(x,\vartheta).$$

The dose *x** is then used to treat the next cohort (or next subject if a cohort's size is one).

A sequential design allows for updating the information (parameter's distribution) after data from each successive subject become available. The dose is increased (or decreased) to the level indicated to be the BOD. This is very similar to the continual reassessment model of (O'Quigley, et al., 1990).To the best of our knowledge, the dose range is usually chosen to be a discrete set of pre-specified values. However, it might be more efficient to allow it to be a real interval, although this may have implications for pharmaceutical formulation. This is another issue for further studies.

The results depend on the initial parameters $\vartheta$, so, to make the decision more robust to wrong specification of the parameters, a Bayesian approach may be used, where some knowledge on the possible distribution of the parameters can be utilised.

The sequential adaptive design is particularly useful when patients rather than healthy volunteers are the first humans treated by a novel agent. This is because this method tends to allocate optimum doses to more subjects and so they may receive maximal benefit.

The trial stops after treating a pre-specified number of patients, provided that at least a minimum required number of patients are treated with the recommended dose level, or a maximum possible (available) number of patients are treated, whichever comes first.

Here we should also mention the importance of numerical simulations of a plausible trial. Various possible scenarios of models, parameter distributions, criteria of optimality can be simulated and so some possible outcomes, such as a number of patients needed for the trial, number of patients treated with an optimum dose, can be predicted to some extent. Then the choice of the best scenario would depend on all the information gathered from pre-clinical studies and of course on the existing knowledge on the novel agent.

### *7.4.3. Allocation of ordinal doses to cohorts*

#### *7.4.3.1. Cohort effects*

Whether or not the dose steps are chosen sequentially, part of the design is to decide how many subjects should be recruited in each cohort, and what proportion of those should receive the available doses or placebo. If the drug is not a novel agent, and does not target a novel part of the body, and has been able to be tested for safety in comparable animals, then one can design a study in which a whole cohort is dosed on the same day or on nearby days.

If, for example, the tolerability of a given dose is assessed, then a crucial choice has to be made. Should the subjects at the given dose be compared only to those subjects given placebo during the same step or to all subjects given placebo? To decide this, we have to ask, is there an effect of cohorts? In other words, might one cohort differ from another in ways that have nothing to do with the different doses? Since different cohorts are treated at different times (planned intervals of at least 14 days in the TeGenero study), a cohort effect may be called a *period* effect. However, cohorts may differ in their subject characteristics as well as in their time of treatment and so we prefer *cohort effect*.

Since TeGenero's Clinical Trial Protocol (page 49) explicitly says that all the placebo data will be pooled, it appears that they had ruled out fitting any cohort effect.
If volunteers are not randomised to cohorts, there will be a cohort effect caused by the different types or availabilities of the volunteers. For example, manual workers are more available in the winter, school teachers in the summer.

Even if volunteers are randomised to cohorts, there may still be a cohort effect. Page 155 of "The textbook of pharmaceutical medicine" recommends that "groups of subjects … be studied together" on "scheduled study days", "thereby expediting the study and enabling efficient use of staff and laboratories"'. This implies that any change in staff (whether clinical or technical), in clinical protocol, in laboratory procedure, let alone the weather, will be confounded with cohorts, and it has been shown over and over again that these effects do produce differences, no matter how much the research organization does its best to ensure that everything is uniform.

The evidence for period effects is backed up by the case studies in "The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales" by (Hu et al., 2005). Although the field is different, the situation may be the same: the medical researchers had no idea that period would have an effect, and the result was that treatment differences were either confounded with, or swamped by, period effects. In one example, all the differences were attributable to six different run dates, rather than to five types of cancer; in another, the major split in the data was not between the two types of cancer but between the measurements made before and after a change in protocol.

Senn gives an example from clinical trials (Section 20.2.5 of (Senn, 1997b)). He asked the physicians if there would be a period effect. They convinced him that there would not be and so they designed for no period effect. The study was delayed and eventually coincided with part of the hay fever season: since one outcome was respiratory function, they had to allow for a period effect in the analysis after all!

#### *7.4.3.2. Suitability of the proposed design*

A justification of the proposed design based on statistical arguments is necessary. For example, the choice of drug/placebo ratio as 3/1 is not explained in the Protocol. This ratio might be viewed as inefficient if the object of the trial is to compare doses with each other or with the placebo. On the other hand, if the aim of the trial is to learn about the safety of a new drug, then we need to get as much information as possible about the PK parameters and adverse events and so there may be little need for placebos.

If cohort effects are fitted, then the subjects at each dose are compared only to those subjects given placebo during the same step. Then the proposed design is inefficient. The variance of any contrast will be proportional to the sum of the reciprocals of the numbers of subjects on each

arm and thus be proportional to 1/6 + 1/2 = 2/3, whereas allocation of four to TGN1412 and four to placebo would have a variance proportional to 1/4 + 1/4 = 1/2, and thus forms a more efficient design. Indeed, a design that employed only six subjects at each step but allocated them in equal numbers to TGN1412 and placebo would have a variance equal to 1/3 + 1/3 = 2/3. In other words, for two fewer subjects per cohort, the same precision could be achieved as for the design actually employed.

The problems with allocation ratios apply *a fortiori* to comparisons of doses. If one wishes to compare two doses in a way that eliminates possible cohort effects then this has to be done using a double contrast whereby each is compared to its own accompanying placebo first. The fact that very few subjects are given placebo in each step constitutes a weak link in the chain of inference: the resulting contrast, for the design employed, has a variance proportional to 2/3 + 2/3 = 4/3. On the other hand, if four out of eight subjects in each cohort receive the placebo, then the variance for the comparison between two doses is proportional to 1/2 + 1/2 = 1.

This means that, if there is a cohort effect, then the design used for the TeGenero study was not *admissible* (Glonek & Solomon, 2004) in the sense that everything that needed to be estimated could have been done, from the same resources, with smaller variance, by a different design (half placebo at each step). In Section 7.4.3.4 we show that even the second design can be improved upon.

However, (Parexel, 2005)(p. 49) stated that the intention was to compare the subjects on a given dose to all subjects given placebo. This immediately raises two issues. First, it implies that formal statistical analysis is not envisaged to guide the decision as to whether to proceed to the next step, since the data from all the subjects constituting the placebo group will not be available until the trial is complete. Secondly, it implicitly accepts the possibility of some bias in comparisons, since the principle of concurrent control is abandoned. In other words, the data were to be analysed on the assumption that there would be no cohort effect.

Under this assumption, subjects in a given cohort are directly compared not only to those in the same cohort but also to those in other cohorts. In fact, this raises another problem with blinding, in addition to those discussed above. The trial is of a form that has been referred to as *veiled* (Senn, 1995), that is to say that although subjects do not know which of five treatments (four doses plus placebo) they are getting at any stage they definitely know three of the treatments they are not getting (the three doses not involved in that step). If, for example, subjects in step 4 believe they are more likely to suffer a problem of tolerability than do subjects in step 1, then this psychological bias can be formally eliminated only by comparing subjects given the highest dose to those given placebo at the same time.

#### *7.4.3.3. Bias-variance trade-offs*

Of course, it may be decided to accept some potential bias and not formally eliminate cohort effects. In fact, if the object is to minimise mean square error in estimation then it is logical to accept some bias as the price of reducing the variance. For example, if all subjects at a given dose are compared to all those on placebo then six are compared to eight and this (potentially biased) comparison has a variance proportional to 1/6 + 1/8 = 7/24, which is smaller than the variance for either design discussed in Section 7.4.3.2 if cohort effects are fitted. This reduction might be considered adequate compensation for any possible bias, may have been the intention in designing this protocol but the fact that this was not made explicit is then regrettable. Also, there remains the problem with analysis at each dose step in order to proceed further.

The fact that the design implies that a bias-variance trade-off is envisaged raises the wider issue as to how far such trade-offs are permissible, for example by comparing results from current subject to historical data.

#### *7.4.3.4. Other possible designs*

The two designs compared in Section 7.4.3.2 are by no means the only possible designs for dose escalation. In order to give a flavour of the wide range of possibilities, but without becoming too technical, in this section we compare some designs for three doses, with eight subjects in each cohort. When there are few concerns about safety, such a design can be used in a conventional way, with the doses chosen in advance and all subjects in each cohort dosed on the same day. The design can also be used if doses are chosen sequentially, as in Section 7.4.2. Then a proper interval is left between dosing subjects within a cohort, and all data from all previous cohorts are used to determine the next dose.

We call the doses 1, 2 and 3. These could be the doses 0.1, 0.5 and 2.0 in the TeGenero study; they could be doses on a linear scale, such as 0.5, 1.0 and 1.5, or on a doubling scale, such as 1, 2 and 4. They could simply be unknown ordinal doses, with only dose 1 known in advance and all subsequent doses chosen sequentially. All that matters for this discussion is that dose 1 is less than dose 2 and dose 2 is less than dose 3. For brevity, we denote the placebo as 0.

The following analysis assumes that outcomes are continuous variables whose variances do not change with dose, but makes no assumptions about the sizes of the differences of effects. A longer, but worthwhile, research project would make similar comparisons for continuous outcomes whose variance is proportional to the mean, and for binary outcomes, such as adverse events.

Table 7.3 displays three possible designs. The third and fourth columns of the table give the variances of the estimates of the differences between doses, and between each dose and placebo. For the first two designs, these are the variances that have been discussed in the preceding text. For the third design, the calculation of variances is more complicated, and we do not go into details here.

Design 1 consists of the first three cohorts of the design used in the TeGenero trial. It is superficially similar to half of the first design shown on page 167 of *The Textbook of Pharmaceutical Medicine*; the difference is that Design 1 uses 32 different human subjects whereas the one in the textbook uses only 8 subjects, using each of them on each of four occasions. The principle is that the *i*-th cohort uses just dose *i* and the placebo, with proportions chosen to ensure that the total number of subjects on the placebo is (approximately) the same as the total number on each dose.

Design 2 is similar to the second design discussed above, which was proposed by Senn (Section 20.2.5 of *Statistical Issues in Drug Development*) (Senn, 1997b). The subjects in cohort *i* are equally split between dose *i* and the placebo. If cohort effects are fitted, Design 2 gives lower variances than Design 1 for all simple contrasts. If there is no cohort effect, and this is known in advance so that the data can be analysed appropriately, then some contrasts have a bigger variance with Design 2 than with Design 1.

Design 3 is from a new class of designs. The principle is that half of the subjects in cohort *i* have dose *i*, the remainder being split between placebo and doses up to *i*-1 in the same proportions as in cohort *i*-1. Obviously, these proportions can be only approximately correct unless the number

of subjects is a suitably large power of 2. If cohort effects are fitted then Design 3 gives lower variances than Design 1 for all simple contrasts; compared with Design 2 it gives lower variances for all but one of the simple contrasts, where it is slightly bigger. If there is no cohort effect, Design 3 is intermediate between Design 1 and 2.

If cohort effects are expected then Design 3 is a better choice than Design 1 or Design 2. If it is not known whether cohort effects will be large, Design 3 is still a good choice, because it is not the worst for either method of analysis. In general, a design is good if it performs at least moderately well over a range of values of the cohort effects(Bailey, 1999). Moreover, Design 3 performs better than the other two designs in terms of blinding, because subjects in cohort 2 have three possible treatments (two doses and placebo) and those in cohort 3 have all four.

The blinding advantage may be diminished if lower doses in each cohort are always administered before higher doses. Then, however, Design 3 gains in terms of more subjects being treated with dose *i* before dose *i*+1 is administered.

**Table 7.3**. **Three designs for three escalating doses**.
The doses are represented by 1, 2, 3 in increasing order, but are not necessarily on a linear or a log scale. The placebo is represented by 0. Each design uses 24 human subjects, 8 in each of 3 cohorts, each cohort being treated and observed before the next starts. In the variance tables, the variance of the estimator of the difference between doses *i* and *j* is the product of the *ij*-entry and the variance per observation. For example, if the variance per observation is $\sigma^2$ and cohort effects are fitted, then the variance of the estimator of the difference between doses 2 and 3 is $1.33\sigma^2$ in Design 1, $\sigma^2$ in Design 2 and $0.72\sigma^2$ in Design 3.

Design 1

| Dose | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Cohort 1 | 2 | 6 | 0 | 0 |
| Cohort 2 | 2 | 0 | 6 | 0 |
| Cohort 3 | 2 | 0 | 0 | 6 |

Variance of differences between doses, and between placebo and each dose, if a cohort effect is fitted

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.67 | 0.67 | 0.67 |
| 1 | | 1.33 | 1.33 |
| 2 | | | 1.33 |

Variance of differences between doses, and between placebo and each dose, if it is known that there is no cohort effect

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.33 | 0.33 | 0.33 |
| 1 | | 0.33 | 0.33 |
| 2 | | | 0.33 |

Design 2

| Dose | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Cohort 1 | 4 | 4 | 0 | 0 |
| Cohort 2 | 4 | 0 | 4 | 0 |
| Cohort 3 | 4 | 0 | 0 | 4 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.50 | 0.50 | 0.50 |
| 1 | | 1.00 | 1.00 |
| 2 | | | 1.00 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.33 | 0.33 | 0.33 |
| 1 | | 0.50 | 0.50 |
| 2 | | | 0.50 |

Design 3

| Dose | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Cohort 1 | 4 | 4 | 0 | 0 |
| Cohort 2 | 2 | 2 | 4 | 0 |
| Cohort 3 | 1 | 1 | 2 | 4 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.29 | 0.40 | 0.65 |
| 1 | | 0.40 | 0.65 |
| 2 | | | 0.58 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.29 | 0.31 | 0.39 |
| 1 | | 0.31 | 0.39 |
| 2 | | | 0.42 |

We are not claiming that Design 3 is the uniquely best design. There is a whole branch of statistics devoted to finding so-called **optimal** designs ((Atkinson & Donev, 1992; Shah & Sinha, 1989). We are just pointing out that a few different (relatively simple) designs could, and should, be compared before a trial is conducted. The protocol should justify why one design was chosen rather than another.

**Table 7.4**. **Four designs for three escalating doses**.
The doses are represented by 1, 2, 3 in increasing order, but are not necessarily on a linear or a log scale. The placebo is represented by 0. Each design uses 32 human subjects, 8 in each of 4 cohorts, each cohort being treated and observed before the next starts. In the variance tables, the variance of the estimator of the difference between doses *i* and *j* is the product of the *ij*-entry and the variance per observation. For example, if the variance per observation is $\sigma^2$ and cohort effects are fitted, then the variance of the difference between doses 1 and 2 in Design 5 is $0.57\sigma^2$.

**Design 4**

| Dose | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Cohort 1 | 2 | 6 | 0 | 0 |
| Cohort 2 | 2 | 2 | 4 | 0 |
| Cohort 3 | 2 | 0 | 4 | 2 |
| Cohort 4 | 2 | 0 | 0 | 6 |

Variance of differences between doses, and between placebo and each dose, if a cohort effect is fitted

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.37 | 0.30 | 0.37 |
| 1 | | 0.47 | 0.67 |
| 2 | | | 0.47 |

Variance of differences between doses, and between placebo and each dose, if it is known that there is no cohort effect

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.25 | 0.25 | 0.25 |
| 1 | | 0.25 | 0.25 |
| 2 | | | 0.25 |

**Design 5**

| Dose | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Cohort 1 | 2 | 6 | 0 | 0 |
| Cohort 2 | 2 | 0 | 6 | 0 |
| Cohort 3 | 2 | 0 | 0 | 6 |
| Cohort 4 | 2 | 2 | 2 | 2 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.36 | 0.36 | 0.36 |
| 1 | | 0.57 | 0.57 |
| 2 | | | 0.57 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.25 | 0.25 | 0.25 |
| 1 | | 0.25 | 0.25 |
| 2 | | | 0.25 |

**Design 6**

| Dose | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Cohort 1 | 4 | 4 | 0 | 0 |
| Cohort 2 | 4 | 0 | 4 | 0 |
| Cohort 3 | 4 | 0 | 0 | 4 |
| Cohort 4 | 2 | 2 | 2 | 2 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.30 | 0.30 | 0.30 |
| 1 | | 0.50 | 0.50 |
| 2 | | | 0.50 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.24 | 0.24 | 0.24 |
| 1 | | 0.33 | 0.33 |
| 2 | | | 0.33 |

**Design 7**

| Dose | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Cohort 1 | 4 | 4 | 0 | 0 |
| Cohort 2 | 2 | 2 | 4 | 0 |
| Cohort 3 | 1 | 1 | 2 | 4 |
| Cohort 4 | 1 | 1 | 2 | 4 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.25 | 0.31 | 0.38 |
| 1 | | 0.31 | 0.38 |
| 2 | | | 0.31 |

| | 1 | 2 | 3 |
|---|---|---|---|
| 0 | 0.25 | 0.25 | 0.25 |
| 1 | | 0.25 | 0.25 |
| 2 | | | 0.25 |

Suppose that we can afford to use an extra cohort, so that there are 32 human subjects in all, but we still have a trial escalating doses from 1 to 3. Table 7.4 compares four possible designs.

Design 4 is based on a crossover design currently used by pharmaceutical companies. The placebo can be inserted into the rising sequence of three doses in any one of four positions. In the crossover design, each of these four sequences is allocated to two subjects. Design 4 is not a crossover design, but the numbers of subjects on each dose in cohort *i* is the same as in period *i* of the crossover design. One advantage is that only two subjects are exposed to the highest dose the first time that it is used. A disadvantage is that doses 1 and 3 are never compared directly, so the variance of the estimator of this difference is comparatively high.

Designs 5 and 6 consist of Designs 1 and 2 supplemented by a fourth cohort, which is equally divided among the three doses and placebo. Thus the unequal replication of Design 2 is preserved in Design 6, so this is the worst of the four designs if there is no cohort effect.

Design 7 repeats the last cohort of Design 3, thus making the design equireplicate. Table 7.4 shows that Design 7 appears to be the best of these four designs if a cohort effect is fitted, and to be no worse than any of the others otherwise.

## *7.5 Conclusions and recommendations*

It seems clear that, contrary to the stated aims of the protocol, the design of the trial of TGN1412 was not suited for testing safety. It may or may not have been suitable for testing tolerability. Even for that purpose there are some aspects of design that could be criticised. However, where the consequences in terms of human safety are not serious it is probably best to avoid being too proscriptive, as investigators who are close to the problems should be given the freedom (within reasonable limits) to choose and justify the solutions they consider best. Nevertheless, the following recommendations are made:

7.5.1. A careful justification of the initial dose should be provided, based on suitable pre-clinical studies.

7.5.2. A risk assessment for the first volunteer should be given.

7.5.3. For genuine first-in-man studies, simultaneous treatment of subjects is inappropriate. A **proper interval** needs to be proposed and observed. This would allow early evidence of toxicity to halt the trial without risk to further subjects.

7.5.4. The proper interval should be defined and receive justification and discussion in the clinical trial protocol.

7.5.5. The plan for blood sampling and analysis and for observation of vital signs should be carefully considered based on the information from pre-clinical studies.

7.5.6. The protocol should specify not only when samples should be drawn but also the time profile for when specific results (such as T cells, IL2, etc.) on those samples will be reported back to the investigators; that is, the likely time delay before laboratory monitoring data are actually available. This matters scientifically (in regard to the "proper interval"), clinically and ethically.

7.5.7. The protocol should specify the method of choosing the second and higher doses, in particular, whether they are based on the pre-clinical information or will be determined sequentially using outcome information from lower doses.

7.5.8. The clinical trial protocol should also discuss the particular design chosen and its limitations and its implications for analysis. For example, if an unequal allocation between treatment and placebo per dose step is chosen this affects the ability of the data-safety monitoring board to have a more efficient assessment of tolerability before proceeding to a further step of the trial.

7.5.9. The clinical trial protocol should allow a protocol review by MHRA and should describe the process in sufficient detail (for example, how far off site is the laboratory which analyses the samples?) to allow an ethical review board and volunteers themselves to determine if the availability of individual subject results (not just the schedule for taking samples) is compatible with contemporary monitoring for the safety of individual participants.

7.5.10. The clinical trial protocol should describe the proposed statistical analysis in sufficient detail to allow an ethical review board to determine if the design is compatible with the proposed analysis.

7.5.11. Both the design of the trial and the analysis of the data should be based on realistic models of the pharmacokinetic behaviour.

7.5.12. In view of the relative rapidity of phase I studies we could consider making it mandatory that a full Statistical Analysis Plan be available as part of the protocol.

7.5.13. For the purpose of evaluating (usually) rare serious side effects, concurrent control is not necessary but background information on the frequency of such events may be extremely useful. To this purpose it would be beneficial for the pharmaceutical industry, either directly or via the regulator to collaborate in building up a shared data-base of adverse events in phase I trials. This requires an agreed protocol for database assembly, for instigating ad hoc studies, as well as a plan and a responsibility for routine regular statistical analyses. There is a link here to Bayesian approaches to quality control using background evidence (Grieve, 1994; Wright, 1992).

## Appendix 1. RSS Working Party Members

Dr Dipti Amin, Senior Vice-President, Global Medical & Regulatory Affairs, Quintiles Limited

Professor R.A. Bailey, Professor of Statistics, Queen Mary, University of London

Professor Sheila M. Bird FFPH, Principal Scientist/Statistician, MRC Biostatistics Unit, visiting professor at Department of Statistics and Modelling Science, University of Strathclyde

Dr Barbara Bogacka, Reader in Probability and Statistics, Queen Mary, University of London

Mr Peter Colman[+], Senior Consultant Statistician, Pfizer Global R+D, Statistical Applications

Dr Andrew Garrett, Vice President Biostatistics, Quintiles Limited

Professor Andrew Grieve, Professor of Medical Statistics, King's College London

Professor Sir Peter Lachmann, FRS, FMedSci, Emeritus Professor of Immunology, University of Cambridge

Professor Stephen Senn*, Professor of Statistics, University of Glasgow

* Chairman
[+] Representative of Statisticians in the Pharmaceutical Industry (PSI)

## Appendix 2. References


Altman, D. G., Gore, S. M., Gardner, M. J., & Pocock, S. J. (1983), Statistical guidelines for contributors to medical journals *British Medical Journal (Clin Res Ed)*, **286**, 1489-1493.

Atkinson, A. C., & Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford: Oxford University Press

Bailey, R. A. (1981), A Unified Approach to Design of Experiments *Journal of the Royal Statistical Society Series A-Statistics in Society*, **144**, 214-223.

Bailey, R. A. (1999), Choosing designs for nested blocks *Listy Biometryczne*, **36**, 85-126.

Berry, D. A., & Eick, S. G. (1995), Adaptive Assignment Versus Balanced Randomization in Clinical Trials: a Decision Analysis *Statistics in Medicine*, **14**, 231-246.

Beyersdorf, N., Hanke, T., Kerkau, T., & Hunig, T. (2006), CD28 superagonists put a break on autoimmunity by preferentially activating CD4+CD25+ regulatory T cells *Autoimmunity Reviews*, **5**, 40-45.

Bretz, F., Pinheiro, J. C., & Branson, M. (2005), Combining multiple comparisons and modeling techniques in dose-response studies *Biometrics*, **61**, 738-748.

Buyse, M., & Molenberghs, G. (1998), Criteria for the validation of surrogate endpoints in randomized experiments [published erratum appears in Biometrics 2000 Mar;56(1):324] *Biometrics*, **54**, 1014-1029.

Calne, R. (2006). *The Daily Telegraph*. Human drug testing (letter to editor), 21 March 2006

CDISC Clinical Data Interchange Standards Consortium. (2006), *CDISC homepage* http://www.cdisc.org/ Accessed 22 October 2006

Colaco, C. (2006 ), What went horribly wrong in a London clinical trial? *The Scientist* **20**, 14.

Dennehy, K. M., Elias, F., Zeder-Lutz, G., Ding, X., Altschuh, D., Luhder, F., & Hunig, T. (2006), Cutting edge: monovalency of CD28 maintains the antigen dependence of T cell costimulatory responses *The Journal of Immunology*, **176**, 5725-5729.

Dhaun, N., Maxwell, S. R. J., & Webb, D. J. (2006). *Behind the Medical Headlines*. Rofecoxib, selective COX-2 inhibitors and cardiovascular risk,

Editorial. (2006a), Institute of Medicine urges reforms at FDA *Lancet*, **368**, 1211.

Editorial (2006b). *The Times*. The wrong dose. Depite the tragey, human testing of drugs is essential, 18 March 2006

Enterprise and Industry Directorate-General. (2006). *Detailed guidance on the collection, verification and presentation of adverse reaction reports arising from clinical trials on medicinal products*
*for human use. As required by Article 18 of Directive 2001/20/EC*, European Commission

EnterTrials. (2006), *FAQs- Learn more about medical trials and volunteering* http://www.entertrials.co.uk/faqs-drug-trials Accessed 9 October 2006

Expert Scientific Group on Clinical Trials. (2006a). *Final report*. London

Expert Scientific Group on Clinical Trials. (2006b). *Interim report and supporting papers*. London

Farrington, C. P., & Whitaker, H. J. (2006), Semiparametric analysis of case series data (with discussion) *Journal of the Royal Statistical Society Series C-Applied Statistics*, **55**, 1-28.

Faunce, T. A., & Bolsin, S. N. (2005), Fiduciary disclosure of medical mistakes: the duty to promptly notify patients of adverse health care events *Journal of Law and Medicine*, **12**, 478-482.

Festing, M. F. W., Overend, P., Das, R. G., Borja, M. C., & Berdoy, M. (2002). *The Design of Animal Experiments*. London: Royal Society of Medicine Press

Food and Drug Administration. (2002). *Guidance for Industry: Estimating the Safe Starting Dose in Clinical Trials for Therapeutics in Adult Healthy Volunteers*, Center for Drug Evaluation and Research (CDER), & Center for Biologics Evaluation and Research (CBER)

Food and Drug Administration. (2006). *FDA and the Critical Path Instititute announce Predictive Safety Testing Consortium will share tests to Understand Safety of Potential New Drugs Earlier*,
Garrett, A. D. (2006).*The Role of Subgroups and Sub-Populations in Drug Development and Drug Regulation*, Statistics, Open University
Glonek, G. F. V., & Solomon, P. J. (2004), Factorial and time course designs for cDNA microarray experiments *Biostatistics*, **5**, 89-111.
Goodyear, M. (2006), Learning from the TGN1412 trial *British Medical Journal*, **332**, 677-678.
Gore, S. M. (1994), The consumer principle of randomisation *Lancet*, **343**, 58.
Grieve, A. P. (1994), A Further Note on Sampling to Locate Rare Defectives with Strong Prior Evidence *Biometrika*, **81**, 787-789.
Gunnell, D., & Ashby, D. (2004), Antidepressants and suicide: what is the balance of benefit and harm? *British Medical Journal*, **329**, 34-38.
Hanley, J. A., & Lippman-Hand, A. (1983), If nothing goes wrong, is everything all right? Interpreting zero numerators *JAMA*, **249**, 1743-1745.
Hawkes, N. (2006a). *The Times*. Drug trial ignored guideline on safety, 16 March
Hawkes, N. (2006b). *The Times*. Stricken drug trial victims face long road to recovery, 18 March
Healy, D., Herxheimer, A., & Menkes, D. B. (2006), Antidepressants and Violence: Problems at the Interface of Medicine and Law *PLoS Medicine*, **3**
Hu, J., Coombes, K., Morris, J., & Baggerly, K. (2005), Importance of Experimental Design in Proteomic Mass Spectrometry Experiments: Some Cautionary Tales *Briefings in Genomics and Proteomics*, **3**, 322-331.
Hutton, J. L. (2000), Ethics of medical research in developing countries: the role of international codes of conduct *Statistical Methods in Medical Research*, **9**, 185-206.
International Conference on Harmonisation. (1995). *ICH Topic E 2 A Clinical Safety Data Management: Definitions and Standards for Expedited Reporting*: European Medicines Agency
International Conference on Harmonisation. (1999), Statistical principles for clinical trials (ICH E9) *Statistics in Medicine*, **18**, 1905-1942.
International Conference on Harmonisation. (2000). *ICH Harmonised Tripartite Guideline: Safety Pharmacology Studies For Human Pharmaceuticals S7a*,
Ioannidis, J. P., Evans, S. J., Gotzsche, P. C., O'Neill, R. T., Altman, D. G., Schulz, K., & Moher, D. (2004), Better reporting of harms in randomized trials: an extension of the CONSORT statement *Annals of Internal Medicine*, **141**, 781-788.
Irving, R. (2006). *The Times*. Drug developers 'will have to carry out test on live organs', 21 March 2006
Jones, G., & Gordon, A. (2006). *The Sunday Express*. Clue that told me to ban Elephant Man drug tests, 19 March
Kadane, J. (1996). *Bayesian Method and Ethics in Clinical Trials Design*. New York: Wiley
Kenter, M. J., & Cohen, A. F. (2006), Establishing risk of human experimentation with drugs: lessons from TGN1412 *Lancet*, **368**, 1387-1391.
Kramer, B. S., et al. (2006), Getting It Right: Being Smarter about Clinical Trials *PLoS Medicine*, **3**, e144.
Laurence, D. J. (2005), Participants in research *British Medical Journal*, **331**, 110.
Laurence, D. J. (2006), No plain English for UK subjects *Applied Clinical Trials*, **15**, 94.
Lee, A. (2006). *The Daily Express*. Do 'guinea pigs' know the risks?, 21 March
Lister, S. (2006a). *The Times*. Drug trial axed after patients poisoned, 15 March
Lister, S. (2006b). *Times Online*. Scientists are cleared of blame for drug trial that went wrong, 6 April
London, A. J. (2006), Reasonable risks in clinical research: A critique and a proposal for the Integrative Approach *Statistics in Medicine*, **25**, 2869-2885.

Mandeville, K. (2006), My life as a guinea pig *British Medical Journal*, **332**, 735-735.
Mayor, S. (2006), Severe adverse reactions prompt call for trial design changes *British Medical Journal*, **332**, 683.
McKie, R., & Revill, J. (2006). *The Observer*. Trial & Terror, 19 March
MHRA Press office. (2006), *Clinical trial suspension: latest findings* http://www.mhra.gov.uk/home/idcplg?IdcService=SS_GET_PAGE&use%20Secondary=true&ssDocName=CON2023515&ssTargetNodeId=389 Accessed 16 December
Modi, M. (2006), Dose-finding studies in phase I and estimation of maximally tolerated doses in *Dose Finding in Drug Development*, (ed. N Ting), pp. 30-48. Springer.
O'Hagan, A. (1998), Eliciting expert beliefs in substantial practical applications *Journal of the Royal Statistical Society Series D-the Statistician*, **47**, 21-35.
O'Neill, R. T. (2002), Regulatory perspectives on data monitoring *Statistics in Medicine*, **21**, 2831-2842.
O'Quigley, J., Pepe, M., & Fisher, L. (1990), Continual Reassessment Method : a Practical Design for Phase-1 Clinical Trials in Cancer *Biometrics*, **46**, 33-48.
Orme, M., Harry, J., Routledge, P., & Hobson, S. (1989), Healthy volunteer studies in Great Britain: the results of a survey into 12 months activity in this field *British Journal of Clinical Pharmacology*, **27**, 125-133.
Parexel. (2005). *Clinical Trial Protocol 68419 : A phase-I Single-Centre Double-Blind, Randomised Placebo-Controlled, Single Escalating-dose Study to Asses the Safety, Pharmacokinetics, Pharmacodynamics and Immunigenicity of TGN1412 Adminstered Intravenously to Healthy Volunteers*. London: Parexel
Peto, R. (1994), Smoking and death: the past 40 years and the next 40 *British Medical Journal*, **309**, 937-939.
Pocock, S., et al. (1991), Statistics and Statisticians in Drug Regulation in the United Kingdom *Journal of the Royal Statistical Society Series A-Statistics in Society*, **154**, 413-419.
Posner. (2005), Exploratory Development in *The Textbook of Pharmaceutical Medicine*, (eds. J.P. Griffin, & J. O'Grady), pp. 144-175. London: BMJ Books.
Prentice, R. L. (1989), Surrogate endpoints in clinical trials: definition and operational criteria *Statistics in Medicine*, **8**, 431-440.
Revill, J., McKie, R., & Hill, A. (2006). *The Observer*. Drug chief defends tests on volunteers. 'Devastated' scientist assumed no side-effects, 19 March
Rogers, L., & Oakshot, I. (2006). *The Sunday Times*. Earlier trial suggested drug in test horror could be toxic, 19 March
Rogers, L., Woods, R., & Deer, B. (2006). *The Sunday Times*. Poison chalice, 19 March
Rosenberger, W. F., & Lachin, J. M. (2002). *Randomization in Clinical Trials: Theory and Practice*. New York: Wiley
Royle, J. M., & Snell, E. S. (1986), Medical research on normal volunteers *British Journal of Clinical Pharmacology*, **21**, 548-549.
Schneider, C. K., Kalinke, U., & Lower, J. (2006), TGN1412--a regulator's perspective *Nature Biotechnology*, **24**, 493-496.
Senn, S. J. (1995), A personal view of some controversies in allocating treatment to patients in clinical trials [see comments] *Statistics in Medicine*, **14**, 2661-2674.
Senn, S. J. (1997a), Are placebo run ins justified? *British Medical Journal*, **314**, 1191-1193.
Senn, S. J. (1997b). *Statistical Issues in Drug Development*. Chichester: John Wiley
Senn, S. J. (2001), The Misunderstood Placebo *Applied Clinical Trials*, **10**, 40-46.
Senn, S. J. (2002), Ethical considerations concerning treatment allocation in drug development trials *Statistical Methods in Medical Research*, **11**, 403-411.
Senn, S. J. (2006), Comment on Farrington and Whitaker: Semiparametric analysis of case series data *Journal of the Royal Statistical Society Series C - Applied Statistics*, **55**, 581-583.

Shah, K. R., & Sinha, B. K. (1989). *Theory of optimal designs*. Berlin: Spinger
Sibille, M., Deigat, N., Janin, A., Kirkesseli, S., & Durand, D. V. (1998), Adverse events in phase-I studies: a report in 1015 healthy volunteers *European Journal of Clinical Pharmacology*, **54**, 13-20.
Simpson, D., & Lee, S. (2003), Tobacco: public perceptions and the role of the industry *Journal of the Royal Statistical Society Series A-Statistics in Society*, **166**, 233-239.
Smith, A. F. M. (1996), Mad cows and ecstasy: chance and choice in an evidence based society *Journal of the Royal Statistical Society Series A-Statistics in Society*, **159**, 367-383.
Stanski, D. R., Rowland, M., & Sheiner, L. B. (2005), Getting the Dose Right: Report from the Tenth European Federation of Pharmaceutical Sciences (EUFEPS) conference on optimizing drug development *Journal of Pharmacokinetics and Pharmacodynamics*, **32**, 199-211.
Strom, B. L. (1995), What is pharmacoepidemiology? in *Pharmacoepidemiology*, (ed. B. L. Strom), pp. 3-13. New York: Wiley.
Suntharalingam, G., Perry, M. R., Ward, S., Brett, S. J., Castello-Cortes, A., Brunner, M. D., & Panoskaltsis, N. (2006), Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412 *New England Journal of Medicine*, **355**, 1018-1028.
The Academy of Medical Sciences. (2005). *Safer Medicines: A report from the Academy's FORUM with industry*. London: Academy of Medical Sciences
The Cochrane Collaboration. (2006), *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5* http://www.cochrane.org/resources/handbook Accessed 26 July 2006
TOPS. (2006), *TOPS: The overvolunteering protection service* http://www.tops.org.uk/site/cms/ Accessed 9 October 2006
Trochim, W. M., & Cappelleri, J. C. (1992), Cutoff assignment strategies for enhancing randomized clinical trials *Controlled Clinical Trials*, **13**, 190-212.
Watts, G. (2003). *The Trouble with Risk*. London: Royal Statistical Society
Whittington, C. J., Kendall, T., Fonagy, P., Cottrell, D., Cotgrove, A., & Boddington, E. (2004), Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data *Lancet*, **363**, 1341-1345.
Wilkie, D. (1997), Mutuality and solidarity: assessing risks and sharing losses *Philosophical Transactions of the Royal Society of London B Biol Sci*, **352**, 1039-1044.
Williamson, P., Hutton, J. L., Bliss, J., Blunt, J., Campbell, M. J., & Nicholson, R. (2000), Statistical review by research ethics committees *Journal of the Royal Statistical Society Series A-Statistics in Society*, **163**, 5-13.
Wood, A. J. J., & Darbyshire, J. (2006), Injury to research volunteers - The clinical-research nightmare *New England Journal of Medicine*, **354**, 1869-1871.
Wright, T. (1992), A Note on Sampling to Locate Rare Defectives with Strong Prior Evidence *Biometrika*, **79**, 685-691.
Wurzner, R., Fernie, B. A., Jones, A. M., Lachmann, P. J., & Hobart, M. J. (1995), Molecular basis of the complement C7 M/N polymorphism. A neutral amino acid substitution outside the epitope of the allospecific monoclonal antibody WU 4-15 *The Journal of Immunology*, **154**, 4813-4819.
Zelen, M. (1992), Randomised consent trials *Lancet*, **340**, 375.
Zhang, W., Sargent, D. J., & Mandrekar, S. (2006), An adaptive dose-finding design incorporating both toxicity and efficacy *Statistics in Medicine*, **25**, 2365-2383.