CancerGrid data standards for the design and analysis of breast cancer clinical trials - Requirements

Peter Maccallum, Steve Harris, January 2007

Commissioning document for NEAT, tAnGo and neo-tAnGo common data elements

Background

Trials models and electronic information systems

The CancerGrid project (www.cancergrid.org) is pioneering the development of advanced clinical trials data management systems. It is using an approach based on standards-based protocol representations conforming to the CONSORT standard [Moher 2001], and registries of re-usable data reporting instruments. This will enable effective electronic information management, long-term re-usability of datasets, and effective integration between clinical and laboratory based outcomes analysis.

Metadata registries and common data elements

The basis of the CancerGrid system is the *common data element*, a full description of a single (simple or complex) item of data which is recorded [Harris 2006]. The common data element is entered into a *metadata registry*, where information about the units of measure (days, millimetres, degrees Celsius), the data type (number, text, choice from a list), instructions on how the data is obtained and what it is taken to represent are recorded. Detailed meanings, choices of terms, and relations between similar data elements from different sources are represented by cross-references with *knowledge services*, which may include clinical terminologies, specialist technical ontologies, or local data dictionaries.

The CancerGrid data standards review process

The common data element approach is based on the architecture of the US National Cancer Institute's Cancer Biomedical Informatics Grid (caBIG) project [Covitz 2003]. In the CancerGrid system, the choice and design of common data elements are to be managed by the *data community*, the study design team and the end users of the data, rather than the informatics specialists who build and maintain the systems but may not have the specialist expertise required to evaluate the issues surrounding data representations.

To support this, the project has built its own metadata registry system, and is constructing an initial common data element subset to represent the clinical trials data which will be analysed – patient outcomes data from three phase III trials.

Scope

Exemplar trials

The CancerGrid project will be providing systems based on three recent or current breast cancer phase III trials

NEAT

The NEAT trial [NEAT 2001] is the National Breast Cancer Study of Epirubicin plus CMF versus Classical CMF Adjuvant Therapy. It is a large scale trial designed to test the following hypothesis: *In women with early breast cancer, adjuvant combination chemotherapy which schedules 4 cycles of epirubicin, followed by 4 cycles of classical CMF, is significantly superior to classical CMF for 6 cycles, in terms of disease-free and overall survival.*

NEAT closed in July 2001, with a sample size of 2000 patients.

tAnGo

The tAnGo trial [TANGO 2004] is a Phase III randomised trial of gemcitabine in paclitaxel-containing, epirubicin based adjuvant chemotherapy for women with early stage breast cancer. It tests the hypothesis: *In women with early stage breast cancer, the addition of gemcitabine to paclitaxel-containing, epirubicin-based, adjuvant chemotherapy provides significantly superior disease-free and overall survival, without excess toxicity or prolonged adverse impact on quality of life.*

tAnGo closed in November 2004, with a sample size of 3000.

Neo-tAnGo

The Neo-tAnGo trial [NEOTANGO 2006] is a neoadjuvant study of sequential epirubicin + cyclophosphamide and paclitaxel +/- gemcitabine in the treatment of high risk early breast cancer with molecular profiling, proteomics and candidate gene analysis. It therefore parallels the tAnGo study in a neoadjuvant setting; neoadjuvant chemotherapy has become standard practice in poor risk, early breast cancer where it may be possible to reduce the extent of surgical treatment and carry out breast-conserving surgery.

Neo-tAnGo opened in 2005, with a target sample size of 800.

Expected uses for the data

The CancerGrid approach is designed to provide flexible, reconfigurable services which can be re-used for different applications. These uses affect the common data elements required. Example applications which are under development include:

Clinical trials management demonstration

A full electronic data capture and data management environment for a trial from start up through to long-term follow up will be built based on the standard protocol model. Fully working examples of NEAT, tAnGo and neo-tAnGo will be demonstrated.

Clinical trials data analysis

Tools to look at the existing clinical outcomes data from NEAT, tAnGo and neo-tAnGo and integrate them with standard statistical analysis workflows will be provided for use by trial investigators and statisticians.

Meta-analysis demonstration

Since the exemplar trials form a related family with similar eligibility criteria and outcomes, and are being subjected to similar tissue-based analysis, the data sets should provide some opportunities for meta-analysis, and the benefits of using a standard set of common data elements to support this will be investigated.

Tissue-based research

Blood and tissue samples from the trial subjects are being collected and will be subjected to analysis at a variety of sites. Detailed pathology and laboratory results are beyond the scope of this initial common data element set, but key to allowing long term integration is to have good management of sample and patient identifiers; common data elements for initial sample collection and tracking are required.

Detailed requirements

The CancerGrid system will require all of the data elements to be captured, but for this detailed review a subset outlined below has been chosen. Since the project will not be gathering data directly, all of this information must be available from existing forms and information systems for the trials.

Administrative information

Common data elements to record patient identity, geographical location of treatment, and identities of staff recording information are required.

Primary endpoints

Common data elements sufficient to represent the information required to analyse the stated primary endpoints for each of the trials are required.

Secondary endpoints

Common data elements sufficient to represent the information required to analyse the stated secondary endpoints for each of the trials are required.

Eligibility criteria

Common data elements to describe the eligibility criteria, common between the trials to allow similarities and differences to be recorded, are required.

Stratification variables

For statistical analysis, common data elements to distinguish all of the stratification variables used in randomisation are required.

Treatment protocol

The treatment protocols used in each arm of the trial need to be represented. Current versions of the CancerGrid model do not include detailed treatment protocol models, but different treatment arms need to be adequately represented. Broader requirements for treatment representation are topics for discussion.

Tissue management

One of the key goals of the CancerGrid project is to provide tools to link clinical outcomes with tissue- and blood-based laboratory analyses. While full pathology and genotyping data representation is beyond the scope of the proposed subset of common data elements, detailed definition of information used to record sample collection and tissue tracking is required.

Review committee composition

The common data element review meeting is scheduled for 16 January 2007. Reviewers will be drawn from the named investigators and management of the three exemplar trials, including

- Clinical co-ordinators
- Translational research coordinators
- Statisticians
- Trial co-ordinators

The results of the review will be recorded in detail, and the final set of common data elements will be published for wider use.

References

[Covitz 2003] Covitz PA, Hartel F, Schaefer C, de Corondao S, Fragoso G, Sahni H, Gustafson S, Buetow KH, caCORE: A common infrastructure for cancer informatics, *Bioinformatics* 19 2404-2412 (2003)

[Harris 2006] Harris S, CancerGrid Report MRC-1.1.3. Common Data Element Representation.

[Moher 2001] Moher D, Schulz KF, Altman DG, Revised recommendations for improving the quality of reports of parallel group randomized trials 2001, *The Lancet* 357 1191-1194 (2001)

[NEAT 2001] Poole C, Earl H, NEAT National Breast Cancer Study of Epirubicin plus CMF versus Classical CMF Adjuvant Therapy, http://www.ncrn.org.uk/portfolio/dbase.asp, accessed 9 May 2006

[NEOTANGO 2006] Earl H, Neo-tAnGo A neoadjuvant study of sequential epirubicin + cyclophosphamide and paclitaxel +/- gemcitabine in the treatment of high risk early breast cancer with molecular profiling, proteomics and candidate gene analysis, http://www.ncrn.org.uk/portfolio/dbase.asp, accessed 9 May 2006

[TANGO 2004] Poole C, tAnGo A Phase III randomised trial of gemcitabine in paclitaxel-containing, epirubicin based adjuvant chemotherapy for women with early stage breast cancer, http://www.ncrn.org.uk/portfolio/dbase.asp, accessed 9 May 2006