Accelerating Cancer Research Using Semantics-Driven Technology

James Brenton*, Jim Davies†, Jeremy Gibbons†, and Steve Harris†
* *Cambridge Research Institute, Cancer Research UK*
† *Oxford University Computing Laboratory*

Abstract

We are embarking on a project to increase the rate of progress in clinical cancer research by accelerating the development, running, and re-analysis of translational early-phase studies through the provision of semantics-driven software solutions. We are taking an approach that has proven effective for randomised controlled trials, where therapies are tested on large populations, and will apply it to the earlier stages of research and development, where the effect of therapies upon humans are first explored and significant discoveries are made. We are building upon the work of the UK CancerGrid project, which has developed techniques for the provision of software systems in which data is associated with a detailed, computable representation of its semantics, and that association is used to drive subsequent processing—analysis, integration, and re-use. This position paper describes the research questions and the technological challenges that arise in addressing these questions.

1 Early-phase studies in cancer

Current chemotherapy strategies for most cancers are empirical; they have had only limited effect in curing patients, because the molecular basis for drug response and resistance is not well understood. This inability to individualize care by choosing the best drug treatments for each patient remains the fundamental clinical problem in oncology. It is therefore essential to develop clinical biomarkers that identify patients who will not respond to chemotherapy, prior to or very soon after initial chemotherapy treatment, so that alternative treatments can be evaluated.

This requires the development of complex early-phase clinical studies that use molecular and functional imaging profiles to better understand the mechanisms of drug action. These phase I or II studies are typically designed to recruit 50–100 patients who are intensively studied; the designs are very different from phase III studies, in which efficacy questions are asked against the current standard treatment in large numbers of patients.

Our work in this area has focused on ovarian cancer, and we have correlated clinical response to carboplatin and palitaxel chemotherapy with molecular profiles of cancer samples collected before and during treatment [2]. These analyses have discovered new and clinically relevant markers for drug resistance, which will be taken forward into new early-phase studies. Our current clinical studies incorporate functional imaging of response in tumour masses using novel magnetic resonance imaging techniques. This is important, as the microenvironment around cancer cells may also have important effects on drug resistance: tumours have variable blood supply, and poorly perfused regions of tumours are hypoxic, which limits delivery and efficacy of anticancer drugs. The novel imaging techniques include *dynamic contrast enhanced magnetic resonance imaging*, *diffusion weighted imaging*, and *magnetic resonance spectroscopy*. Our current studies incorporate multiple tissue sampling of tumour masses from areas of high and low perfusion. By combining the analysis of longitudinal changes in these imaging parameters with changes in regional molecular profiles, we are likely to discover better response biomarkers.

2 Clinical research challenges

There are significant challenges in developing, running and analysing early-phase studies such as the above: the scientific questions to be addressed by the study may need complex designs and sample collections; authoring, reviewing and implementing study protocols is time consuming, requires considerable input from clinical investigators and clinical trial staff, and provides very

limited scope for reuse of artefacts such as designs and data collection forms from previous studies; studies may need to be carried out at multiple centres, as only a small number of patients may be suitable (particularly where molecular criteria are needed for entry), and investigators need to have large numbers of studies open at each site; data sharing of image and molecular data from these studies is extremely difficult.

The UK CancerGrid Project [1] has previously focused on supporting the design and running of phase III clinical studies, together with the creation of well-annotated tissue resources for prognostic marker studies. This has led to the development of a metamodel [3] for phase III trial design, and tools for protocol authoring using controlled vocabularies and common data elements. These models will need to be extended to the very different requirements for early-phase work. In particular, this will require detailed methods for specifying tissue collection, linkage with standard operating procedures, and different endpoints and designs.

Extending the CancerGrid work to phase I/II studies has key advantages: the cost and risk of phase I/II studies is several orders of magnitude less than a phase III trial; success in this area will encourage the rapid adoption of CancerGrid technologies in high-profile academic medical research facilities; and reuse of data and analysis will accelerate clinically relevant discoveries in cancer care—a greater proportion of early-phase studies will yield re-usable results, and the time taken to develop and deploy a complex study protocol will be reduced (based on our experience with phase III trials, we may see a reduction of up to 80%).

3 Technological research challenges

The approach taken by the CancerGrid project is *semantics-driven*: careful attention is paid to the collection and curation of semantic terminologies; trial designs are annotated with semantic metadata; and the model-driven technology for generating trial support software maintains the association between data and metadata by construction. There are four aspects of the current approach that will require significant extension to enable the modelling and operation of early-phase clinical studies, outlined below.

3.1 Metamodelling marker studies data and acquisition processes

The CancerGrid trials model is based upon CONSORT [4], a standard for the reporting of randomised controlled trials with a well-defined hypothesis, and a clear conception of how to measure efficacy. The principal analysis is declared in the study protocol before the first patient is recruited, and will enshrine a fixed conception of success. Patients in a typical phase III trial will not have received prior treatment for their disease, and the experimental treatment will be a viable, ethical option for patients meeting the eligibility criteria.

Early-phase studies are quite different in nature, with prior treatment an important consideration, along with the identification of toxicity and side effects, and indicators of drug action: for example, the rate at which the drug is broken down by the metabolism, and the extent to which specific tissues are retaining or reacting to the drug. These studies may undergo considerable evolution during execution, to the point at which every single patient may be subject to a different treatment regime. In an early-phase study, we expect a wider variety of data items to be considered, recording detailed information about a small number of patients. More importantly, we expect the nature of the study to change significantly as it progresses, adapting to take account of information obtained.

3.2 Federation of metadata registries

The present approach, in which collaborators make use of a single metadata registry in the design and execution of a trial, does not scale well to support the kind of loosely-knit, multidisciplinary collaboration required for sophisticated, early-phase studies. Different communities will wish to (and are best placed to) maintain their own collections of metadata elements, value sets, and experimental designs. On the other hand, a centralised approach, in which different communities own different sections of a single, monolithic registry, lacks the agility and flexibility to support the required degree of scientific innovation and variation of practice across a range of diseases and disciplines. What is needed instead is a distributed approach, in which the focus is upon

versioned collections of metadata elements, rather than upon the registry (or registries) that hold them.

3.3 Semantic integrity of models

The present semantic framework places no constraint upon the way in which metamodels are instantiated with metadata elements to produce study designs. This is tolerable in the development of phase III randomised trials, where there is a single, regimented study design, with clearly defined objectives, developed by a small group of closely collaborating individuals, and applied for the whole duration of the study.

In early-phase studies, in contrast, study designers will need reassurance that the metadata elements selected are being used in an appropriate context: there is greater diversity in study design; the precise objectives may not be clear at the outset; there may be only loose collaboration in design, as a researcher modifying one aspect may not be fully aware of the intentions behind related aspects; and the study design evolves during execution. It is important that designers are guided, or constrained, when selecting metadata elements to represent observations about a particular entity. For example, in designing a class to represent a patient carer, we would not wish to add a metadata element denoting a clinical observation associated with a particular disease. The usage of metadata elements should be consistent with their semantics, as described by attributes or relationships in the metadata registry.

3.4 Analytical techniques and tissue management

The existing CancerGrid metamodel is focussed upon clinical data and the processes surrounding its collection. It does not have support for the description of analytical techniques, or associated administrative activity. This has proved tolerable, thus far, in application to phase III trials, where the experimental activity is declared in advance, easily expressed, and does not change during execution.

An early-phase marker study may involve varying application of analytical and statistical techniques at different points in its execution, and precise nature of the variation must be accurately recorded in the (evolving) study model. This will require metamodelling of scientific workflows, in order that the history of study models will present a sufficiently detailed account of how the results were obtained. This level of detail is required also in respect of physical, chemical, and biological procedures outside the experimental data set: for example, the time elapsed between tissue removal and subsequent freezing, the calibration of instruments, or the precise techniques used in the preparation of tissue products. To assess comparability of observations, we must be able to measure the extent of compliance with standard operating procedures, and relate alternative procedures for the same technique.

References

[1] J. Brenton, C. Caldas, J. Davies, S. Harris, and P. Maccallum. CancerGrid: developing open standards for clinical cancer informatics. In *Proceedings of the UK e-science All Hands Meeting 2005*, pages 678–681, 2005. `http://www.allhands.org.uk/2005/proceedings/`.

[2] H. M. Earl, A. Ahmed, A. Vallier, H. Hatcher, C. Parkinson, M. Iddawela, J. Latimer, R. Crawford, and J. D. Brenton. Expression profiling of advanced epithelial ovarian cancer (EOC) to predict chemotherapy response. *Journal of Clinical Oncology (2006 ASCO Annual Meeting Proceedings)*, 24(18S):15018, 2006. Cambridge Translational Cancer Research Ovarian Study 01 (CTCR-OV01).

[3] S. Harris and R. Calinescu. CancerGrid clinical trials model 1.0. CancerGrid technical report MRC/1.4.1.1, 2006. `http://www.cancergrid.org/public/documents`.

[4] D. Moher, K. F. Schultz, and D. G. Altman. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, 357, April 2001.