Addressing informatics challenges in Translational Research with workflow technology

Simon A. Beaulah[1], Mick A. Correll[2], Robin E.J. Munro[1] and Jonathan G. Sheldon[1]

[1] InforSense Limited, Colet Court, 100 Hammersmith Road, London W6 7JP, UK
[2] InforSense LLC, 155 Second Street, Cambridge, MA 02141, USA

Interest in Translational Research has been growing rapidly in recent years. In this collision of different data, technologies and cultures lie tremendous opportunities for the advancement of science and business for organisations that are able to integrate, analyse and deliver this information effectively to users. Workflow-based integration and analysis systems are becoming recognised as a fast and flexible way to build applications that are tailored to scientific areas, yet are built on a common platform. Workflow systems are allowing organisations to meet the key informatics challenges in Translational Research and improve disease understanding and patient care.

The pharmaceutical and biotech industries continue to struggle against difficult market conditions caused by increasing R&D costs, high late-stage attrition, safety concerns, poor pipelines and pricing pressure from healthcare bodies and generics. This has resulted in decreasing valuations and shareholder confidence, and a poor stock market for life sciences in the past three years, while the rest of the market has grown strongly [1]. This is forcing a dramatic re-think in many parts of the business; one such area of re-invention is making drug discovery more patient-focused and integrated with healthcare provision, as suggested by the NIH Road map for accelerating bench to bedside discoveries [2]. Such Translational Research is seen as a way to lever the knowledge and expertise gained from using ‘omics and other such technologies to increase the understanding of patients, bringing improvements in attrition and delivery of targeted, safe and effective drugs.

Translational Research aims to bring together experimental approaches that are commonly used in drug discovery, such as gene expression, Genome Wide Association (GWA) studies and proteomics, with data from clinical trials and more general patient information [3]. Combining this information and making it easily accessible to researchers enables them to design and test their hypotheses more effectively. By integrating these previously separate domains into a two-way process, Translational Research can improve disease understanding, identify disease and efficacy biomarkers [4] and improve patient care. The resulting vision for drug discovery companies is more effective and safer therapeutics, and a reduction in spiralling drug development costs [5].

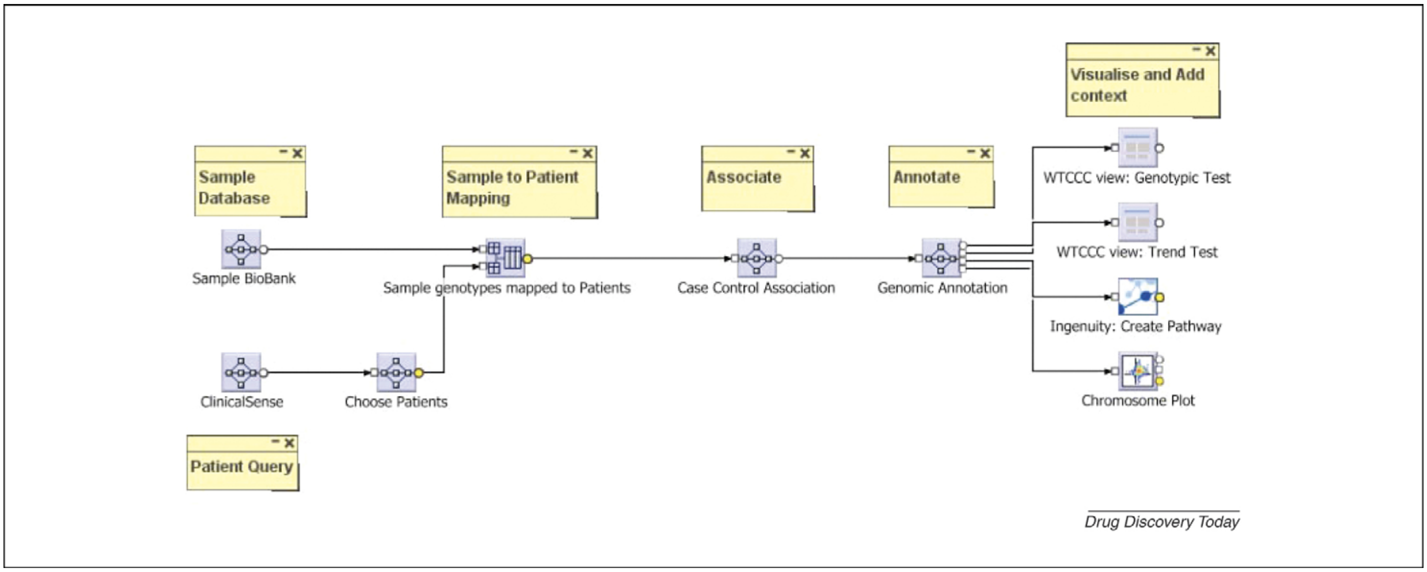
To provide an effective foundation for Translational Research, an informatics infrastructure must support a number of capabilities: advanced data integration techniques to bring together clinical, sample and research data; powerful tools to normalise and analyse translational data; appropriate user interfaces and visualisation tools for technical and scientific user groups and the flexibility to adapt to new techniques, technologies and processes. Bringing clinical and research data together is, therefore, a daunting prospect, but without an effective informatics infrastructure Translational Research is severely impaired.

Workflow technologies

Workflow-based systems allow analytical applications to be constructed visually without the need for coding (Figure 1), based on data manipulation and analysis components, which can be extended to incorporate new data sources and algorithms. This type of visual programming enables non-software developers to build and deliver analytical applications. Workflow systems provide a graphical environment in which data and analysis components can be combined, usually representing a business or scientific process, in an interactive environment. These workflows can then be run by an underlying workflow engine, meaning that the process is not only captured, but also can be run as a

Corresponding author: Sheldon, J.G. (jsheldon@inforsense.com)

1359-6446/06/$ - see front matter © 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2008.06.005



Drug Discovery Today

FIGURE 1

A workflow combining patient and sample data, matching subjects into patient groups and analysing the resulting genotypic data before displaying it in a pathway view and chromosome plot. Boxes represent data sources or analysis steps and the arrows indicate data flow.

programme as well. This provides a powerful integration and analytical environment to build scientific applications rapidly. Workflow systems are widely used to drive web interfaces and enable custom interfaces to be quickly developed and deployed. They therefore lend themselves to providing specific interfaces for different user groups. With commercial tools available from, for example, InforSense (see: http://www.Inforsense.com), and open source tools from Taverna [6], workflow systems are growing in popularity within Life Sciences [7–11].

Workflow systems are particularly suited to addressing the IT challenges inherent in Translational Research (Figure 2). They are able to simplify the delivery of complex science because applications can be constructed and modified quickly. As a result, software teams are very productive and are able to support multiple groups of clinicians and researchers and respond to the rapidly changing requirements found in Translational Research. Statisticians and bio-informaticians use workflows to produce the best practice data analysis and make it available to a wider scientific audience via simple web interfaces, improving consistency. Advanced analytical workflows can automate quality control (QC) and analysis of individual 'omics data sets and then combine them to build predictive disease and healthcare models. Powerful database access capabilities enable workflow tools to integrate clinical, sample and experimental data into analysis workflows. This combination of factors means that Translational Research can be scaled up to address many therapeutic areas without overwhelming statisticians, IT infrastructure and development teams.

The following sections review the data integration, data analysis, user delivery and flexibility challenges of Translational Research in more detail, and illustrate, using case studies, how workflow systems have been used to address them.

Data integration

Biological systems are inherently complex and many scientific fields have developed to study genes, proteins, chemicals and diseases, resulting in an equally complex data and informatics infrastructure. For the pharmaceutical industry, it often means a series of isolated domains of research that each focus on a particular area of science. Communication with other departments can be poor, particularly between clinical and pre-clinical organisations. This disjointed infrastructure is viewed as being unsustainable, and approaches such as Translational Research are driving a more integrated environment [12]. Having recognised these problems the industry is seeking to change culturally, and by developing informatics systems that are able to transcend boundaries. There are significant IT challenges in integrating clinical data from patient histories, lab results, pathology reports and clinical questionnaires [13]. Similar challenges can be found when integrating experimental drug discovery data, where integration of siloed data has long been a thorn in the side of IT departments.

Patient data, when in an electronic format, are often stored in multiple databases and in different formats. It usually has an operational focus that is not appropriate for supporting disease studies. Before patient data can be integrated with research data they need to be cleaned, normalised and combined into a single source, usually an HL7 data warehouse (see: http://www.hl7.org). This can then be used to assess potential patient populations for appropriate conditions defined by the researcher or clinician. For clinical trials and cohort studies, users need to explore and monitor populations of patients in case and control groups on the basis of study criteria. They then check whether tissue or blood samples are available via the LIMS or BioBanking systems [14].

Even with HL7 databases, compliance with protected health and privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA), is essential. An HL7 data warehouse can still be too focused on operational data and access to confidential data is closely restricted. This can be addressed using a disease oriented data mart that has de-identified patient data that can be used as the basis of Translational Research. With pass-through security from the data mart back to the HL7 database it is still possible to access secure patient data with appropriate permissions. For pharmaceutical companies, clinical trials bring the same

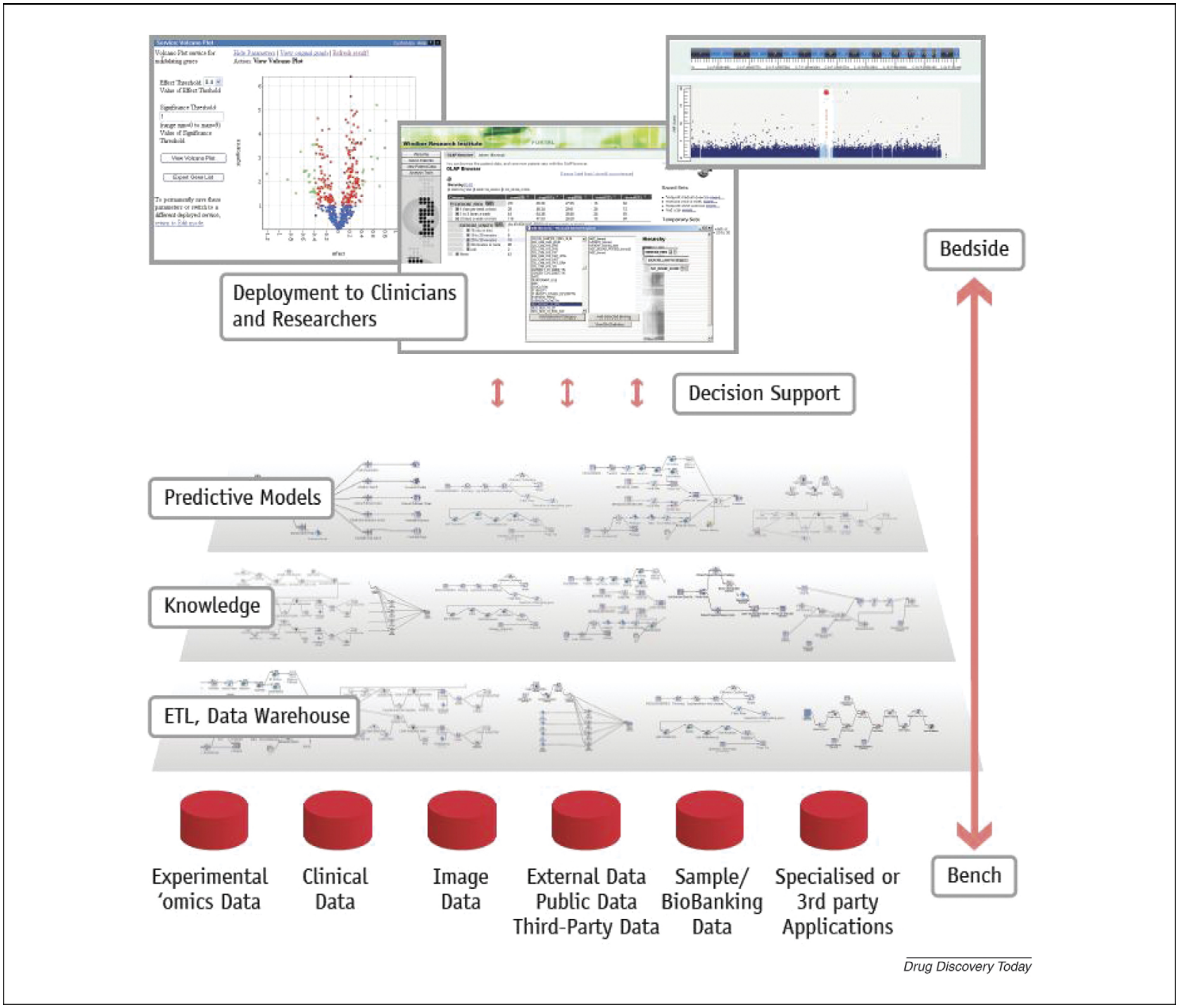


FIGURE 2

An informatics architecture for Translational Research based on workflow technologies. Workflows are used by IT to integrate translational data into the environment and into data warehouses, by statisticians to represent business logic for the QC and analysis of data and to develop predictive models of disease and biomarker prediction, and then by IT to deliver this information to clinicians and researchers via web interfaces.

compliance issues so de-identified versions of clinical data need to be brought into the discovery part of the business.

Service-oriented architectures are becoming common in Life Sciences [15], meaning Translational Research infrastructures need to integrate Web Services programming interfaces as well as databases. Initiatives such as caBIG (see: http://cabig.nci.nih.gov) and Informatics for Integrating Biology and the Bedside (see: http://www.i2b2.org), funded by NIH, provide valuable Web Services that can be used internally by life science companies and medical institutes to annotate and understand data. In addition, access to internal data and tools will increasingly need to be made available via Web Services to support data integration and web-based delivery to end users.

A Translational infrastructure needs to support both data warehousing, when data can be pulled together, and data federation for when sources need to be kept separate. The integration of patient and sample data is a complex process that must be carefully defined and managed, whether it is building a data warehouse or federating data sources. This type of problem is well addressed by workflow technologies that can query individual databases, retrieve relevant data, filter and clean the data and place them in a data mart that supports Translational Research rather than operational review. Where formation of new data marts and warehouses is not an option, workflow systems can federate data and Web Services; for example, to integrate a database of gene expression information using the sample IDs as a common identifier. By developing the extract, transform and load or federation processes in workflows, the process is clearly defined and easily referenced. They are not hidden in code and SQL calls, and can be easily modified and extended. The modular nature of workflows, where each one addresses a different point of the integration, also makes for easier understanding, documentation and maintenance.

Semantic integration of data is also a key requirement in Translational Research. The growing use of ontologies and the Semantic Web promises much in delivering common identifiers and relationships in Life Sciences [16]. Workflow systems can benefit greatly from advances in the Semantic Web by integrating semantic web services directly into workflows to facilitate the combination and analysis of data.

Data integration case study: Dana-Farber cancer institute

The most valuable resources in the study of diseases such as cancer are disease samples collected from patients who are undergoing treatment and on whom relevant clinical and outcome data are available. In pursuit of its mission to defeat cancer, scientists at the Dana-Farber Cancer Institute have amassed a substantial collection of patient samples (with the appropriate informed consent) and a vast body of clinical data associated with the patients contributing those samples. Designing experiments to use these samples to advantage, and analysing the resulting data, requires that scientists have direct access to complex data that are often distributed among various data stores across the Institute. As such, even fundamental questions involved in experimental design such as 'Identify patients with CD138+ bone marrow cells, who are newly diagnosed with multiple myeloma and divide them into groups of responders and non-responders to bortezomib treatment' are difficult to answer. Researchers are required to complete forms and ask technical staff to query databases, which is often takes time to be completed.

To address these questions, a pilot study in multiple myeloma is integrating patient and sample data into an HL7 data warehouse [17]. Workflows are being used to enable scientists to directly access these data to select patient cohorts [18]. The combination allows scientists to link information regarding availability, age and storage medium of samples such as bone marrow, serum and urine. Previously these studies could only be designed through extensive trial and error discussions between the lead scientists and data analysts with access to the data stores. With the benefit of an integrated system, considerable time is saved in defining studies. In addition, subtleties in the data can be identified that previously would have remained hidden owing to the long turnaround times. Once a patient subset has been selected, gene expression experiments are run and data integrated, via the sample ID, with the clinical data used to design the experiment. This integration is absolutely essential for analysis of the data, focusing on identifying genes whose patterns of expression can classify the patients into relevant groups based on response to therapy. Workflows are also being used to provide automated annotation of gene lists from public sites such as GO, UNIPROT and OMIM.

Data analysis

Analysis of scientific data requires many algorithms to identify patterns and relationships and predict outcomes. To support Translational Research, informatics infrastructures must be able to incorporate any algorithm into an analysis environment that connects to the previously described data integration layer. Each high-throughput technique has algorithms for QC and analysis, often producing a gene or protein list. The lists then require validation using metabolic pathway tools or scientific literature searches. The arrival of high-throughput technologies has resulted in new algorithms that are able to identify patterns and are scalable to work in data sets of many millions of data points. Standard statistical tools such as SAS (see: http://www.SAS.com) and R (see: http://cran.r-project.org) are therefore supplemented with technique-specific analysis such as Copy Number Variation [19] and Hardy–Weinberg [20] tests. As well as 'omics technologies, Translational Research frequently includes imaging techniques that also need to be analysed and compared. For example in oncology, assessment of tumour size and growth is driven by image analysis providing good biomarkers of disease status and progress.

In a translational study a number of techniques are often used simultaneously because the combination of different approaches provides a more accurate prediction. For example, a study to identify biomarkers for Alzheimer's disease [21] uses a combination of proteomics, clinical questionnaires, Magnetic Resonance Imaging and lipidomics to identify suitable combinations of factors. This requires the data from each technique to be checked and normalised before being integrated into a summary table using the subject ID. Anova, *t*-tests and principal component analysis are then used to identify predictive biomarkers.

Using predictive modelling to improve clinical decision making is fundamental to Translational Research. Therefore, translation infrastructures need to support these techniques as well as multivariate analysis techniques and statistical analysis extensions. Multiple models need to be run against multiple experimental data sets so that they can be compared and contrasted to find the best approach. This is an iterative process requiring a series of trial and error assessments before deployment to end users. The result provides predictive models to support clinicians and enables Translational Research to make knowledge-driven decision making a reality by putting information into context. Knowledge-driven decision making is new to Translational Research [22] but offers tremendous potential for improved healthcare.

Workflow systems support data analytics and mining capabilities [23], and enable multiple modelling techniques to be run over data sets. The results can then be compared in order to identify the most appropriate modelling approach or data set. Their ability to support the rapid construction of analysis workflows means that the iterative time required to perfect a model is significantly reduced, and service groups are able to respond more quickly to requests from users. Once the workflows are complete they can be made available to researchers and clinicians who are able to adjust particular parameters to test the models.

Data analysis case study: Erasmus MC

Erasmus University Medical Centre (Erasmus MC) provides advanced medical care for a local population of three million people in The Netherlands. Erasmus MC is actively developing evidence-based clinical models for oncology, cardiovascular and neurological Translational Research approaches that integrate clinical, gene expression, imaging, sample and genetic approaches. Analysis of this data is complex and Erasmus MC is using workflows as a basis for their analysis platform [24]. Each technique requires its own QC before analysis; Genome Wide Association studies, in particular, require large quantities of data to be filtered by allele frequency, genomic location and Hardy–Weinberg equilibrium. Once checked, data need to be combined to allow multivariant analysis and predictive models to be run, requiring access to statistical and modelling tools. It is often a case of trial and error to find the best classification or predictive model, so Erasmus MC use workflows to run multiple tools and compare the results. Erasmus MC is seeking to guarantee the best experimental accuracy and reproducibility of experimental results, to ensure that the resulting models are accurate and do not suffer from errors due to the selection process or too limited statistical significance. Workflows are being developed to optimise classification within patient cohort studies using methods such as Decision Trees and Support Vector Machines. Within this framework researchers can explore the relationship between sample size, outcome, methods, validation and parameters for optimal patient stratification in areas of high un-met medical needs.

User interaction

Translational Research requires good collaboration between many scientific and technical groups. Each group requires its own user interfaces and tools to support its particular needs. Without tailored user interfaces, translational projects are likely to be greatly impaired. The following list illustrates the types of group involved in Translational Research projects and their particular needs:

- Clinicians need to interact with patient records and predictive models via simple web pages to support data-driven clinical decision making.
- Biomedical research scientists need to access data and analysis workflows via easy to use web pages built by bio-statisticians and IT personnel.
- Bio-statisticians need to develop analysis workflows rapidly using R and SAS tools; they also require predictive modelling, clustering and classification algorithms.
- IT personnel need a standard-based platform with strong database and web services support, as well as powerful workflow and web deployment tools.

Access to and analysis of data is frequently a bottleneck for researchers and clinicians, as the turnaround time for a new analysis can be a few weeks owing to the heavy workload of statisticians and bioinformatics teams. Scientists need direct access to data and analysis, but following dictates defined by statisticians and in a format that is end user focused and not overwhelmed by complex interfaces. Web pages are the best deployment option for such scientific groups, as they allow users to view and assess data in ways that are tailored to them. Bio-statisticians and IT developers need a powerful desktop tool to construct analysis and integration workflows.

User interaction case study: Windber Research Institute and Walter Reed Army Medical Center

Research and clinical teams at Windber Research Institute and Walter Reed Army Medical Center are using workflow systems to support multiple user groups through the same command-and-control informatics platform [25]. IT personnel have used workflow builder clients to integrate data from multiple clinical sources to provide a unified set of patient data, which has then been combined with tissue sample, genomics, proteomics and imaging data into a single patient-centric model. Bio-statisticians use workflow builder clients to design and test analysis protocols that can then be deployed to researchers and clinicians.

Researchers use a web interface to define cohorts in their translational studies based on patient, sample and experimental data across the organisation. They can see how many patients fit into case or control groups and expand the detail of each patient to minimise selection bias. A longitudinal view of a patient's history also supports this. Researchers can browse and drill into data from hundreds of parameters; for example, a breast cancer study had over 500 different factors that could be assessed. Translational experiments can then be run, and researchers can access the data through the same system and run complex analyses written by bio-statisticians. Unnecessary complexity is hidden and researchers only see a few relevant options and the results.

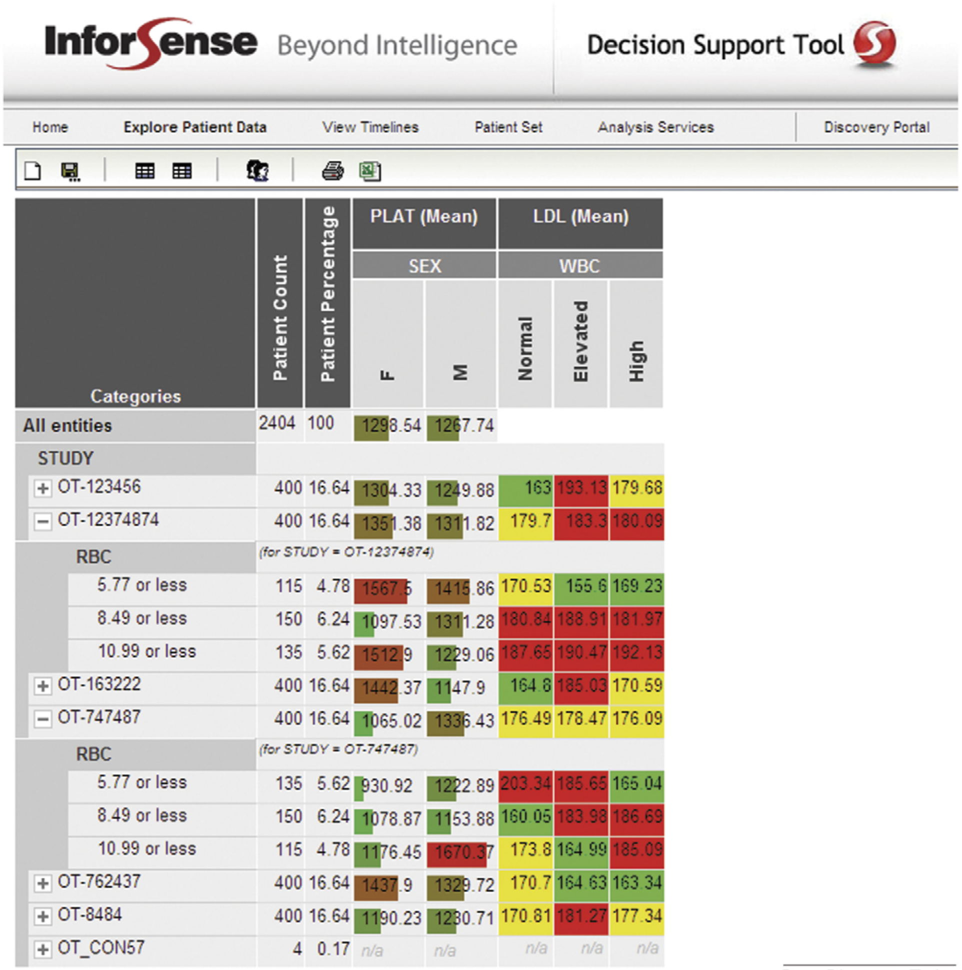
Clinicians also use web interfaces (Figure 3) based on the same system, but modified for their needs, to improve clinical decisions. A population explorer can be used to find past patients who closely match new patients entering treatment. Past patients' treatment and response can then be mapped on a timeline to look for potential risk factors to monitor, and predictive models built by researchers can help to identify the most appropriate treatment. This collaborative system is enabling scientists to translate their research into real decisions that impact patient care, and providing clinicians with the best possible data on which to base decisions.

Flexibility

Translational Research is a new area and the techniques and best practices involved are still evolving. There will always be new algorithms that need to be integrated, such as the CHIAMO SNP testing programme developed by the Wellcome Trust Case Control Consortium [26]. Researchers and statisticians need to evaluate and modify workflows continually to reflect experience gained on past projects and insights gleaned from scientific literature. Disease investigation is often unpredictable and while Translational Research aims to make it more data-driven, it still requires a dynamic environment that supports frequent modifications. Monolithic solutions requiring long development cycles are not going to be effective in Translational Research; infrastructures need to be dynamic to keep pace with changing requirements.

Groups addressing different diseases will need differing techniques and analyses to meet their particular requirements. However, the underlying system requires the same data integration, analysis and deployment capabilities. For example, groups studying endometrial cancer will have different needs to those working on metabolic disorders. IT Systems that have been constructed to integrate clinical and research data for a particular project may not be able to integrate a new technique easily or modify analysis protocols for new areas.

The visual programming approach supported by workflow systems means that once the data sources and algorithms required are integrated, new applications and workflows are easy to create. New workflows and applications can be built and delivered in a few hours, making considerable time savings. Modifying existing workflows is even faster, minimising support and maintenance costs and enabling developers to focus on new and added-value applications.



Categories	Patient Count	Patient Percentage	PLAT (Mean)		LDL (Mean)		
			SEX		WBC		
			F	M	Normal	Elevated	High
All entities	2404	100	1298.54	1267.74			
STUDY							
+ OT-123456	400	16.64	1304.33	1249.88	163	193.13	179.68
− OT-12374874	400	16.64	1351.38	1311.82	179.7	183.3	180.09
RBC	*(for STUDY = OT-12374874)*						
5.77 or less	115	4.78	1567.5	1415.86	170.53	155.6	169.23
8.49 or less	150	6.24	1097.53	1311.28	180.84	188.91	181.97
10.99 or less	135	5.62	1512.9	1229.06	187.65	190.47	192.13
+ OT-163222	400	16.64	1442.37	1147.9	164.8	185.03	170.59
− OT-747487	400	16.64	1065.02	1336.43	176.49	178.47	176.09
RBC	*(for STUDY = OT-747487)*						
5.77 or less	135	5.62	930.92	1222.89	203.34	185.65	165.04
8.49 or less	150	6.24	1078.87	1153.88	160.05	183.98	186.69
10.99 or less	115	4.78	1176.45	1670.37	173.8	164.99	185.09
+ OT-762437	400	16.64	1437.9	1329.72	170.7	164.63	163.34
+ OT-8484	400	16.64	1190.23	1230.71	170.81	181.27	177.34
+ OT_CON57	4	0.17	n/a	n/a	n/a	n/a	n/a

Drug Discovery Today

FIGURE 3

Cohort selection and patient browser web interface developed for clinicians at Windber Research Institute.

Flexibility case study: CIDR at Johns Hopkins University

The Center for Inherited Disease Research at Johns Hopkins University (JHU) is a centralised facility that provides genotyping and statistical genetics services. It is used by JHU investigators seeking to identify genes that contribute to human disease. JHU is active in many areas of research including cancer, diabetes, drug abuse, neurology and stroke. CIDR provides a good example of why Translation Research infrastructures need to be flexible; it has developed a series of workflow tools to monitor quality and produce a variety of reports and data files for Illumina (see: http://www.illumina.com) and Affymetrix (see: http://www.affymetrix.com) GWA arrays [27]. During the QC stage of GWA, researchers from different disease areas need to test alternative QC filters and thresholds. There is no single analytical process that can satisfy every GWA study and therefore each QC workflow needs to be easy to modify. For example, a study into rheumatoid arthritis would commonly exclude single nucleotide polymorphisms (SNPs) from Chromosome 6 so that major histocompatibility complex (MHC) genes are not included. Alternatively a specific genomic region may be of interest in which case other SNPs would be excluded. Workflow systems offer this flexibility in refactoring and allow rapid iteration of analysis processes and re-deployment. They also provide extensibility in cases where users wish to integrate new algorithms, manufacturers and tools in this rapidly changing field.

Conclusion

In this paper we have discussed the complex informatics issues surrounding the support of Translational Research. We have shown how workflow technologies can be used to address the informatics challenges of integrating and analysing data, deploying end user solutions and providing the flexibility required to evolve best practices. The combined strengths of database access capabilities, advanced analytics and deployment options mean that workflow systems can respond quickly to new user requirements and changes in scientific processes. Bringing together combinations of data and techniques in workflow systems enables the promise of Translational Research to be achieved in a repeatable and scalable way. This results in better disease understanding, informative biomarkers and more effective therapeutics for an industry struggling with safety and financial concerns.

Acknowledgements

The authors would like to thank all the partners mentioned in this paper for their groundbreaking work developing the Translational Research systems of the future. Thanks also to the entire InforSense team for providing the technology, skills and dedication to make these projects happen.

References

1 Billion dollar pills. *The Economist* January 25th 2007
2 Zerhouni, E. (2003) The NIH Roadmap. *Science* 302, 63–72
3 O'Connell, D. and Roblin, D. (2006) Translational Research in the pharmaceutical industry: from bench to bedside. *Drug Discov. Today* 11, 833–838
4 (2007) Translational Research in the pharmaceutical industry: from theory to reality. *Drug Discov. Today* 12, 419–425
5 Gilbert, J. *et al.* (2003) Rebuilding big pharma's business model. *In vivo: the Business and Medicine Report.* (vol. 21)
6 Oinn, T. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20, 3045–3054
7 Huang, J. *et al.* (2007) Discovering multiple transcripts of human hepatocytes using massively parallel signature sequencing (MPSS). *BMC Genomics* 8, 207
8 Laaksonen, R.A. (2006) Systems biology strategy reveals biological pathways and plasma biomarker candidates for potentially toxic statin-induced changes in muscle. *PLoS One* 10.1371/journal.pone.0000097 http://www.plosone.org
9 van der Weyden, L. *et al.* (2006) Loss of *TSLC1* causes male infertility due to a defect at the spermatic stage of spermatogenesis. *Mol. Cell Biol.* 26, 3595–3609
10 Rodriguez, A. *et al.* (2007) Effects of iron loading on muscle: genome-wide mRNA expression profiling in the mouse. *BMC Genomics* 8, 379
11 Clemente, E.J. *et al.* (2006) Gene expression study in the juvenile mouse testis: identification of stage-specific molecular pathways during spermatogenesis. *Mamm. Genome* 17, 956–975
12 Gaughan, A. (2006) Bridging the divide: the need for translational informatics. *Pharmacogenomics* 7, 117–122 Online publication date: 1 January 2006
13 Webb, C.P. and Pass, H.I. (2004) Translation research: from accurate diagnosis to appropriate treatment. *J. Transl. Med.* 2, 35
14 Anderlik, M. (2003) Commercial biobanks and genetic research: ethical and legal issues. *Am. J. Pharmacogenomics* 3, 203–215
15 Curcin, V. *et al.* (2005) Web services in the life sciences. *Drug Discov. Today* 10, 865–871
16 Ruttenberg, *et al.* (2007) Advancing translational research with the semantic web. *BMC Bioinformatics* 9 (Suppl. 3),
17 Quackenbush, J. (2008) An integrated framework for multiple myeloma research. In *Proceedings of Bio-IT World Conference*, Boston, MA, United States, April 28–30th
18 InforSense Signs Dana-Farber as Latest Partner for Translational Research Platform, (2008) *Bioinformatics*
19 Redon, R. *et al.* (2006) Global variation in copy number in the human genome. *Nature* 23, 444–454
20 Stern, C. (1943) The Hardy–Weinberg law. *Science* 97, 137–138
21 Lovestone, S. *et al.* (2007) Biomarkers for disease modification trials—the innovative medicines initiative and AddNeuroMed. *J. Nutr. Health Aging* 11, 359–361
22 Liebman, M.N. (2007) Personalized medicine: a perspective on the patient, disease and causal diagnostics. *Personal. Med.* 4, 171–174 Online publication date: 1 May 2007
23 Tiwari, A. and Sekhar, A.K.T. (2007) Workflow based framework for life science informatics. *Comput. Biol. Chem.* 31, 305–319
24 Stubbs, A. (2008) Biomarker discovery for athlerosclerosis. In *Proceedings of 8th European Biomarker Congress*, Manchester, United Kingdom, 14–15th May
25 Hu, H. *et al.* (2004) Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. *Pharmacogenomics* 5, 933–941
26 Wellcome Trust Case Control Consortium, (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 447, 661–678
27 Munro, R. *et al.* (2007) Building workflows for the quality control of genome wide association data. *Poster Presentation at American Society for Human Genetics*, San Diego, USA, 23–27th October

Reviews • INFORMATICS