The ENSAT Registry: A Digital Repository Supporting Adrenal Cancer Research

Abstract The very nature of rare diseases means that information is often sparse and highly distributed, and as a result progress in the field is more challenging to conduct. ENSAT-CANCER is an EU-FP7 funded initiative to develop a virtual research environment (VRE) offering a digitally interconnected infrastructure for distributed clinicians specialising in rare adrenal tumours to communicate and collaborate with distributed biomedical research communities. The core of the VRE is a registry that holds vital patient information from specialist centres around Europe, covering different types of adrenal tumours. The VRE also hosts a range of other enabling services including sample barcoding, bio-sample exchange mechanisms, an integrated linkage scheme to other trials and studies, summary statistics and report generation, image hosting – all of which are available in a seamless, security-driven environment. . This paper presents the key challenges of this endeavour, the technical solutions that have been developed to address them and reporting the uptake and adoption of the work (currently at 2472 patient records and rising).

Keywords: Cancer registry, Digital biobanking

1. Background

The European Network for the Study of Adrenal Tumors (www.ensat.org) was founded in 2002 through the merging of three existing but largely independent adrenal tumor research networks in France, Germany and Italy, with research teams from the UK. The central aim of the ENSAT consortium has been to improve the prediction and management of malignant adrenal tumors. In particular the community have focused on four main tumor types: adrenocortical carcinoma (ACC), pheochromocytoma and paragangliomas (Pheo/PGL), non-aldosterone producing adrenocortical adenoma (NAPACA) and aldosterone-producing adenoma (APA) - all of which are relatively rare (3.5% of the population have *adrenal incidentalomas*, of which only a subset are malignant) and with typically poor survival rates. It is hoped that the study of the genetics and treatment of adrenal tumor patients will reveal new molecular mechanisms of the growth of these tumor types and provide insight into associated areas, e.g. the role of peptides and steroids in hypertension (a common side effect of adrenal tumors). The diversity and often aggressive/fatal nature of adrenal tumors make them an important condition to address. However, their comparative rarity requires many international resources be drawn upon in order to make significant progress in the field. Given this, a long-term goal of ENSAT is to bring together a critical mass of expertise and resources to achieve significant clinical and biological conclusions and to eventually combat adrenal tumors.

Tumor information about these conditions has hitherto been diverse with isolated repositories of highly valuable data invisible to the vast majority of clinicians and researchers. The need for coordinated data sharing is thus compelling. However, a parallel effect of this rarity is to make the identity, and hence security, of the patient involved inherently more complex. As such, competing technological and informatics tensions exist that the registry and associated e-Infrastructure – including potential data linkage outside the consortium – must address directly.

Contemporaneously, a multitude of clinical trials are being conducted with principal investigators that are either, or have close collaborative links with, the main protagonists of the ENSAT-CANCER initiative. Four of these studies are already well under way (ADIUVO [1], FIRSTMAPPP [2], EURINE-ACT [3] and PMT [4]), others are starting to appear. Whilst funded in their own right, coordination of these trials through the ENSAT-CANCER VRE is essential. For instance, prospectively enrolled patients in a particular study may have information useful to the central adrenal research community, whilst any information held in the registry can provide retrospectively enrolled patient data for a given study.

There are a number of other initiatives in this space. Drives to link clinical health data and the amassing of large, representative biomaterial of a certain population are core goals of CaBIG [5], BioGrid [6] and the UK Biobank project [7]. Whilst these projects have a large amount of effort and momentum, it can be argued that their scale and their scope impinge on their overall utility since they are designed to service the needs of many communities. The ENSAT registry is designed to address the needs of a very target audience, who actually drive its development in a tight iterative feedback cycle [8, 9]. Thus the usefulness can already been seen in the record numbers uploaded – 2472 patients already registered across all tumour (ACC, APA, NAPACA, Pheo) types after a year of operation – which, given the rarity of the conditions, is already an enormous success for the project and the community at large. Nevertheless a variety of challenges remain to be tackled.

2. Data Challenges

2.1. Identification, linkage and security

Central to the information model used in the ENSAT-CANCER project is the notion of a unique patient index that encapsulates the patient identity within a centre in a particular country. This identity is critical to maintaining the consistency of the information held within the registry but it must also be decoupled from the local institutional identifier according to the specifications laid out in the (approved) ethics application for the project. Such identity mapping fits with the requirements of the ISO standard (27001) section on identity management [14].

The numbering itself is facilitated through use of a centralised – separate – database that maintains a running count for each centre registered in the VRE. In centres conducting studies that are related to the ENSAT registry – for instance, the PMT study in Dresden (Pheo), and the ADIUVO study in Turin (ACC) – the identification mapping, and the callout to this separate numbering database, has had to be dealt with on a case-by-case basis. In the PMT study, the electronic Case Report Forms (eCRFs) and associated numbering scheme are under the control of the developers responsible for the ENSAT registry, with full access and control to all of the databases involved. Whereas in the ADIUVO clinical trial, a web service connection has been implemented between the registry and the study eCRFs – maintained by a separate set of developers, who work in collaboration with the ENSAT developers to establish this connection and numbering scheme consistency. As the ADIUVO study commenced before the ENSAT-CANCER initiative, the study's own numbering scheme was already in place. The only way to address this disconnect of two unrelated schemes is to maintain an updateable mapping file that relates the generated patient IDs. Similar issues existed in the bulk upload of legacy data, where the mapping of meta-data information was required between French and German centres. With the consensus of the leading clinicians involved, the French model was adopted (an Access database with tables laid out according to a consortium-agreed standard) and the German model (Excel spreadsheet with a less organised format) was modified to fit the tabular layout.

2.2. Transfer Functions

Corollary to the notion of central identification numbers, is the ability to transfer patient data between the research communities involved in the collaboration as accurately and seamlessly as possible, with as small an impact on the clinicians/researchers as possible. The scheme to achieve this was first prototyped and rolled out in production between the ENSAT registry and the PMT study based in Dresden, Germany. This utilised a back-end database schema common to both the PMT and ENSAT-CANCER systems. Due to the differing requirements of the registry and the study, a net effect – acknowledged as acceptable by the clinicians involved on either side of the transfer – was the "dilution" of information as it passes to the registry. For instance, detailed information on biochemical measurements of plasma and urine metanephrine and methoxytyramine levels are critical in the PMT study and are thus highly specified. In the registry the overall outcome of the information – for instance, what the levels indicate clinically – rather than the detailed measurements, is the key information.

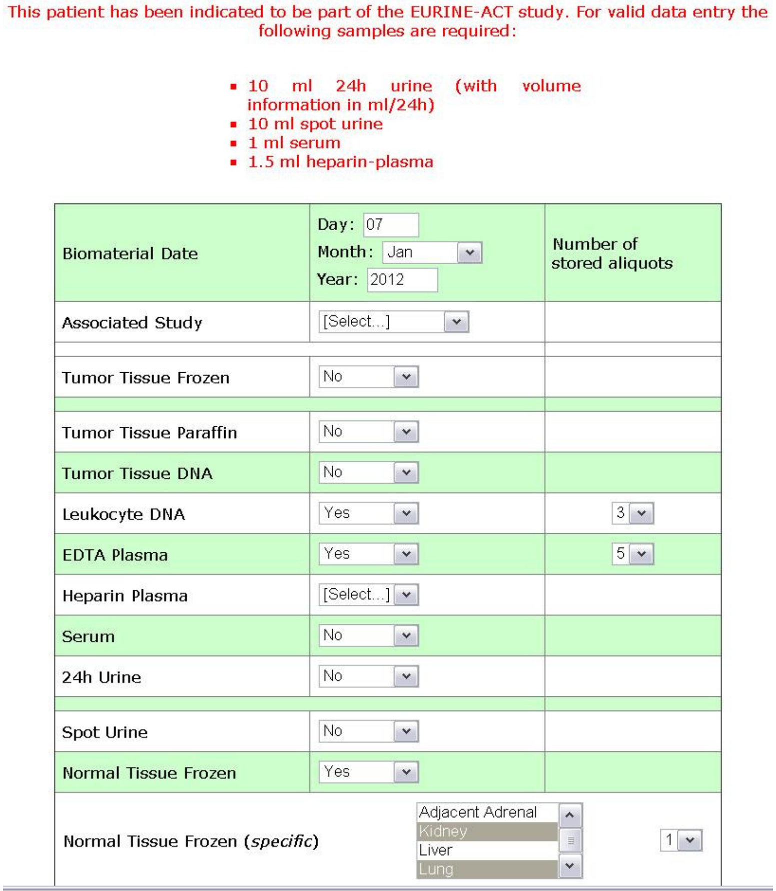
Another feature is the ability for patients to be transferred from the PMT study – a study investigating pheochromocytomas – into the NAPACA section of the registry. This is enabled to allow unspecified adrenal masses – which may yet still present as pheo – to be entered into both the study (before confirmation or exclusion of the pheo) and the registry (as the adrenal mass can be of interest for those focusing on the study of NAPACA-type tumours). These considerations are also present in the transfer function developed between the ADIUVO study and the ACC section of the registry, though the implemented mode of web service differs from the direct database connections used for PMT (more detail is provided in section 3).

2.3. Biobank Features

A prominent role for the ENSAT registry is to provide a centralised biobanking facility that aids the exchange of sample and tissue data across international boundaries. There are core capabilities that have been identified as critical to making this application effective, and these have been developed in consultation with the centres that possess, distributed, process and analyse such data. For instance, the laboratories at the Technische Universität Dresden (Germany) and the University of Birmingham (UK) have been identified as "high-throughput" centres – where many samples are processed using high throughput –omics data processing, e.g. LS/MS mass spectrometers. Other lower throughput data processing centres are also present, such as INSERM (Institut National de la Santé et de la Recherche Médicale) in Paris, France, and the CNIO (Fundacion Centro Nacional de Investigaciones Oncologicas Carlos III) in Madrid, Spain.

To standardise the biomaterial exchange, a core set of parameters have been laid out in forms that belong to a patient record (figure 1). Since many of these forms can be required per patient, clinicians are strongly encouraged to associate each form with a particular study (“ADIUVO”, “PMT”, etc). When that study association has been made, the required items are pre-populated into the form to validate the form data. For instance, for the EURINE-ACT study, specific levels of urine, plasma and serum are required (also shown in Figure 1).

A further critical component to the system is the ability to print labels that will then be affixed to the tissue sample tubes before they are shipped between centres. The information provided includes a high-level description of the contents, the relevant identification numbers, the aliquot number (where one sample is divided into aliquots for logistical purposes), and a barcode to allow ease of scanning by high-throughput machines. There are many barcode standards to choose from – QR codes were initially tried but rejected due to the two-dimensional nature being unreadable on the curved surface of a typical sample tube. As a result the project has now adopted Interleaved 2 of 5 [8] as the agreed standard.



This patient has been indicated to be part of the EURINE-ACT study. For valid data entry the following samples are required:

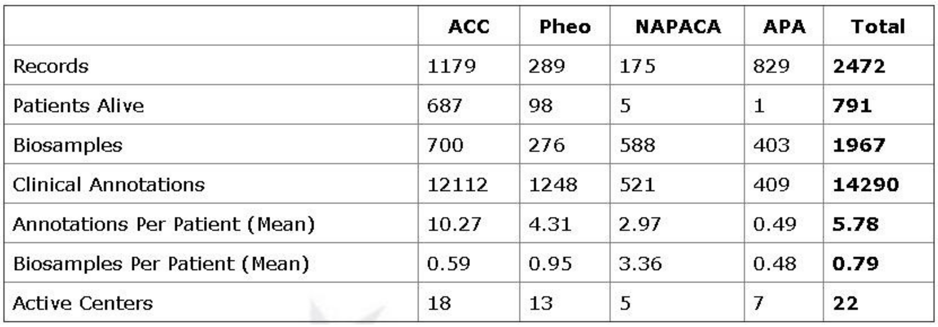
- 10 ml 24h urine (with volume information in ml/24h)
- 10 ml spot urine
- 1 ml serum
- 1.5 ml heparin-plasma

Biomaterial Date	Day: 07 Month: Jan Year: 2012	Number of stored aliquots
Associated Study	[Select...]	
Tumor Tissue Frozen	No	
Tumor Tissue Paraffin	No	
Tumor Tissue DNA	No	
Leukocyte DNA	Yes	3
EDTA Plasma	Yes	5
Heparin Plasma	[Select...]	
Serum	No	
24h Urine	No	
Spot Urine	No	
Normal Tissue Frozen	Yes	
Normal Tissue Frozen (*specific*)	Adjacent Adrenal Kidney Liver Lung	1

Figure 1: standard set of biomaterial parameters required for sample/tissue exchange.

3. Implementation

The VRE has been implemented using well-supported open-source technologies that underpin a stable and robust platform upon which the various services can be reliably provisioned. Figure 2 shows a snapshot of recent record summary.

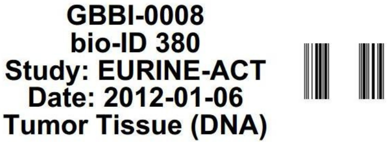


	ACC	Pheo	NAPACA	APA	Total
Records	1179	289	175	829	**2472**
Patients Alive	687	98	5	1	**791**
Biosamples	700	276	588	403	**1967**
Clinical Annotations	12112	1248	521	409	**14290**
Annotations Per Patient (Mean)	10.27	4.31	2.97	0.49	**5.78**
Biosamples Per Patient (Mean)	0.59	0.95	3.36	0.48	**0.79**
Active Centers	18	13	5	7	**22**

Figure 2: a snapshot of record numbers and biomaterial/clinical annotations as of the 20th February 2012.

The registry has been developed using JSP server-side scripting hosted in a Tomcat container, running on a virtualized host (supported by the NeCTAR cloud program at the University of Melbourne). This system provides secure access to a MySQL database holding the primary datasets for the four distinct components of the ENSAT registry. The current features supported include the standard create, read, update and delete functions of repositories. Also provided are search, export to CSV file format, PDF format labels (available for individual printing or in an A4 sheet – this is done using iTextPDF [9]), and summary statistics of clinically relevant items (e.g. a status report that shows in five lines the most important information for cancer follow-up: the condition of the patient, their survival time, time without disease or time to recurrence). To maintain the identification number, a central "ID clock" database – also MySQL-based – is maintained that is interrogated whenever a new patient record is generated. This database is also available to the PMT and ADIUVO studies – directly through a JDBC connection integrated with the PMT record generation code. The connection with the ADIUVO trial makes use of a RESTful [10] web service exposing selected elements of the trial database eCRFs. It is considered acceptable to return the ADIUVO ID (note, this is not the local institution ID) and, when assigned an ENSAT ID based on the next value stored in the ID clock for that centre, is maintained in a mapping file to associate the two IDs.

Finally, the barcodes in the label printing are generated using the zxing Google code project [11], which supports many standards (QR, ITF (Interleaved 2 of 5), Codabar, etc). Figure 3 shows an example of such a label, with all the relevant information and the bio-form ID encoded next to the text.



GBBI-0008
bio-ID 380
Study: EURINE-ACT
Date: 2012-01-06
Tumor Tissue (DNA)

Figure 3: label with sample information and Interleaved 2 of 5 encoded ID

In terms of security, the registry itself is served over https, thereby providing an encrypted connection, and a certificate is distributed to users along with instructions on how to load this into the browser (often a non-trivial task for the average clinical user). The platforms upon which the application is built – database, server, virtual machine – have all administrative options locked down as is best practice for any such service (remote login using SSH keypair, root login disabled, restricted access to database and web-app container, etc). And the application itself has programmatic controls in place to further minimise potential intrusion vectors (rejecting SQL injection attacks, unfamiliar byte-code characters, boundary/sanity-checking of all form input data). Nightly backups are enabled using cron-jobs and mysqldump scripts, which are then transferred to a separate machine. Once a month, these are encrypted using TrueCrypt and sent to the off-site backup store in Munich, Germany (to the lead investigator of the ENSAT-CANCER project).

4. Conclusion and Acknowledgements

The ENSAT-CANCER initiative has provided a highly functional and useful digital repository for the storage and communication of information on rare adrenal tumours. The VRE supports a multitude of functions that visibly and effectively support scientific progress by allowing researchers to conduct international studies that would otherwise be impossible to undertake. The features supported have been implemented with a mind to future-proofing the exchange of information and data for adrenal cancer research. Key future work includes encouraging greater uptake of the bio-banking facility, implementation of a image hosting service outlined in a later work-package of the project, importing new data-sets from other smaller country-specific repositories, e.g. the German CONN and Cushing databases, and supporting large data export projects such as the Stage III/IV ACC investigation being conducted by the Institut Gustav Roussy in Paris and the AVIS-2 APA project by the Universita degli Studi di Padua.

The authors would like to acknowledge the EU FP7 program (grant agreement number: 259735) for kindly funding the ENSAT-CANCER project.

5. References

[1] ADIUVO - https://www.epiclin.it/adiuvo
[2] FIRSTMAPPP - http://www.pressor.org/information/collaborative-projects.htm
[3] EURINE-ACT - http://www.ensat.org/images/EURINEACTInfo.pdf
[4] PMT – https://pmt-study.pressor.org
[5] CaBIG - https://cabig.nci.nih.gov/
[6] BioGrid - http://thebiogrid.org/
[7] UK Biobank - http://www.ukbiobank.ac.uk/
[8] Stell, Sinnott, Jiang - Enabling secure, distributed collaborations for adrenal tumor research - proceedings of the HealthGrid 2010 conference, Paris, France
[9] Sinnott, Stell – Towards a virtual research environment for international adrenal cancer research – Proceedings of ICCS 2011, Singapore
[10] Interleaved 2 of 5 - http://en.wikipedia.org/wiki/Interleaved_2_of_5
[11] iTextPDF - http://itextpdf.com/
[12] RESTful web services - http://en.wikipedia.org/wiki/Representational_state_transfer
[13] Google ZXing project - http://code.google.com/p/zxing/
[14] ISO 27001 - http://en.wikipedia.org/wiki/ISO_27001