Improving Data Quality in Disease Registries and Clinical Trials: A Case Study from the ENSAT-CANCER Project

1,2 Stephan Glöckner, 2Wiebke Arlt, 2Irina Bancos, 1Anthony Stell, 1Richard O. Sinnott
1 Department of Computing and Information Systems,
University of Melbourne,
Melbourne 3010, Victoria
2 School of Clinical & Experimental Medicine,
University of Birmingham,
Birmingham, B15 2TT, UK
rsinnott@unimelb.edu.au

Abstract
The Internet-era has given rise to a global increase in information availability and coupled with technologies that allow data to be produced at an exponential rate; a maelstrom of data, information and potential for new knowledge now exists. This has huge implications for the clinical and biomedical sciences in the post-genomic era where whole genome and exome sequencing technologies are increasingly prevalent and producing copious amounts of sensitive data at an unprecedented rate. This deluge of information is both a challenge to researchers and to clinical care providers as well as an opportunity. To tackle this, many biomedical research communities focus on subsets of information through establishing targeted disease registries consisting of clinical (phenotypic) information on patient cohorts. These registries are often used for targeted recruitment to clinical trials and studies. One such example of this is the European Network for the Study of Adrenal Tumours (ENSA T – www.ensat.org), which has established an extensive range of deeply phenotyped information on patient cohorts with different adrenal tumour subtypes. This resource is used to support a portfolio of clinical trials covering international phase 1 to phase 4 studies. One of these is the EURINE-ACT study. In this paper we focus specifically on the data quality issues that have arisen in ENSAT and the demands of studies such as EURINE-ACT. We identify future steps that are to be taken to improve the data quality including automated data quality assessment scores and user community feedback assessment.

Keywords: Disease Registries, Clinical Trials, Data Quality.

1 Introduction
Biomedical research and research breakthroughs that are achieved should ideally lead to improvements in clinical care across the population. However there are several hurdles that must be overcome in ensuring that this is the case (Khoury et al. 2007). Firstly, the exponential growth of biomedical findings and web-based information more generally is reaching a tipping point. New approaches and new technologies encompassing the omics landscape through to whole genome and exome sequencing are increasing the body of information at an unprecedented rate. There are multitudes of research avenues being pursued across the biomedical landscape that make raising awareness of specific results especially challenging in the large. To tackle this and ensure that focused results are garnered that can/should be translatable, targeted registries and clinical trials focused on specific problems and/or targeted disease areas that are relevant to a much smaller communities represent one approach to overcome the data deluge. However disease registries and clinical trials depend on data and especially data quality. The increased used of Web-based systems allow capturing data across multiple centres and often across multiple countries, however ensuring the overall quality and completeness of the data is often a challenge (Buchan & Bishop 2009; Prokosch & Ganslandt 2009).

This issue is relevant to all research disciplines but is magnified in the health sciences (Chen et al. 2011). Being able to trust data and the analysis of data is essential and can have major repercussions if ignored (Venet et al. 2012). Because of the increasingly diverse and complex range of medical research areas, translating research results into actual clinical settings is a major challenge. While some of these challenges are based on policies and translation of proven research outcomes, a further challenge is a technological one covering both bioinformatics and medical informatics (Payne et al. 2009; Sarkar 2010; U.S. Department of Health and Human Services & Food and Drug Administration 2005). Whilst bioinformatics provides the tools and algorithms that can be used to support data processing, integration, analysis and visualization across clinical genomics, genomic medicine, pharmacogenomics and genetic epidemiology (Sarkar et al. 2011), the repeatability of bioinformatics results is often challenging. This is exacerbated by the amount of web resources dealing with omics data increasing exponentially and the lack of documented/proven approaches for bioinformatics tool chains that have been shown to work at the level of accuracy demand for transfer of results into a clinical setting. On the other hand, medical informatics can be
used to support a range of tasks that can be used to capture and process phenotypic/clinical data, e.g. electronic Case Report Forms (eCRFs) and patient-based disease registries (Brandt et al. 2006). However again, the abundance of web-based resources makes the derivation of knowledge that can be translated into a clinical environment non-trivial, and importantly the overall quality measures of such resources need to be systematically understood and ideally, empirically measurable. The combination of bioinformatics and medical informatics should decrease the translation turnaround time of new developed medical entities (treatments, drugs, diagnostics) from the lab to patients (Sarkar 2010; Woolf 2008). However over the last years the development of new treatments has decreased, whilst the money spent on research and development of new therapies has actually increased (Woodcock & Woosley 2008). Furthermore, up to now new findings in genomics have yet to show an increase in late phase clinical trials (Clinical Research Society 2014; Green & Guyer 2011). Currently an average drug development costs of the order of US $1 billion and the associated development time takes around 10 years. Only 15% of new therapies even get into phase 3 trials and 50% of the novel therapeutics, that pass this barrier actually get approved (Ledford 2011).

For research to move into clinical settings it is essential that every genomic and associated medical development must be unambiguously demonstrated – typically in a late phase clinical trial to show evidence of the effectiveness and efficacy of the new result (Green & Guyer 2011). Often a comparison trial against the Gold standard test (diagnostic study) or standard of care treatment (therapy trial) vs a new diagnostic tool or a placebo must be conducted (Vaidyanathan 2012). The ability to rapidly support trials and studies and ensure the overall quality of data that is collected is essential in this context. New methods in clinical research must be cheaper, faster, more adaptable and more in tune with genomics (Ledford 2011); clinical information must be observed in relation to genotype data; biomarkers across study populations need to be shown (or not!) that they lead to new stratified study designs or biomarker trials (Jensen et al. 2012). In all of these, primary hurdles to be overcome include improved data validation processes to achieve rapid knowledge transfer between the research environment, the laboratory and ultimately to hospital/clinical care environments and patients (McShane et al. 2013).

To obtain new conclusions for novel medical hypotheses, multi-center clinical trials are often needed, but more often than not they don’t show definitive outcomes (Dechartres et al. 2011) but are sensitive to the studies and the way in which statistics are undertaken. In addition to the benefits of an increased sample size and associated improvements in the statistical power leading to better outcomes, multi-center trials enable heterogeneity of the performance and thereby improve the generalizability of the results (Guthrie et al. 2012). However multi-center clinical trials also face challenges. The heterogeneous performance of trials can be impacted by ongoing changes of the study protocols; by overly complex protocol designs, and importantly the demands for extensive information on patients that has to be of high quality (Brandt et al. 2006). The idealized trial should provide a single unambiguous source of centralized information that allows for all data on the trial to be analysed using appropriate tools and methodologies. Ideally these tools should be developed and used in accordance with Good Clinical Practice (GCP) guidelines and applicable regulations including quality assurance procedures of the results.

Quality Assurance (QA) in a clinical research content is defined as “…all planned and systematic actions that are established to ensure that the study has been performed and the ‘clinical’ data are generated, documented (recorded), and reported in compliance with GCP and applicable regulations.” (International Conference on Harmonization 1996). To ensure these aims, several Quality Control (QC) methods have been developed to assess and improve the quality and integrity of the collected data. One of these methods that assess the quality of trial performances is (Data-) Monitoring (Campbell & Sweatman 2002; International Conference on Harmonization 1996). Data monitoring needs to be an on-going process in QC that continuously aims to decrease inconsistencies and/or errors (Arts et al. 2002). Retrospective data quality checks are often impossible to address errors; missing data may not be possible to capture again, and the conclusions of the studies may be invalidated without care and attention to the data capturing process. Data quality and ensuring quality of clinical data as an ongoing process is thus highly desirable.

It is relatively easy to insist that all clinical data should be of the highest possible quality, however for many research endeavours, the quality of data comes at a cost. What data can/should be collected for particular diseases? How important is some data compared to other data? Should all data be compulsory or can some data be optional? The more extensive the data set to be collected, the more of a challenge that exists in ensuring that all hospitals/sites actually collect this information. The role of standardization and consensus across the research community is key here and pragmatic concerns often need to be adopted.

This paper focuses on one initiative (ENSAT – www.ensat.org) that has established a unique set of data resources (https://registry.ensat.org) that have galvanized an international clinical research community and are currently being used to underpin an extensive portfolio of large-scale international clinical trials. We focus in particular on the lessons being learnt in systematically measuring data quality and processes put in place to improve the overall data quality driven by large-scale international clinical studies.

2 Background to the European Network for the Study of Adrenal Tumours

The European Network for the Study of Adrenal Tumors (ENSAT) 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 was to improve the prediction and
management of specific types of adrenal tumours. In particular the community focused on the tumour types: adrenocortical carcinomas (ACC), pheochromocytoma and paragangliomas (Pheo/PGL), non-aldosterone producing adrenocortical adenoma (NAPACA) and aldosterone-producing adenoma (APA) - all of which are relatively rare (e.g. for ACC this is 1.5:1million) and have typically poor survival rates (Golden et al. 2009; Kebebew et al. 2006). Through establishing a critical mass of information on patients with these adrenal tumour types, ENSAT-CANCER aims to show how a portfolio of coordinated studies and trials of the genetics and treatment of adrenal tumour patients can reveal new molecular mechanisms of the growth of these tumour types and provide insight into associated clinical research areas, e.g. their role in hypertension (Beuschlein 2013).

Building on the ENSAT network and the community of adrenal tumour researchers, the ENSAT-CANCER project was funded as part of the EU Framework 7 initiative in 2011. At the heart of the ENSAT-CANCER project was the establishment and support of an advanced clinical research platform – a security-oriented, web-based virtual research environment (VRE). It was intended that this VRE would support all aspects of the collaboration including collection, sharing and analysis of biomedical data (both physical and digital) associated with adrenal tumours (ACC, APA, Pheo/PGL, NAPACA), and the interactions of the clinical and research communities involved. Already, this VRE has galvanized a community of adrenal tumour research efforts across the EU and indeed globally - far beyond the original ENS@T-CANCER project members. It offers a rich range of functionality supporting the biomedical and clinical communities. At present the VRE is used by 53 major clinical cancer centres from across the world including groups in Europe, South America and Russia; it supports advanced biobanking facilities and is used across over 20 major international clinical trials and studies (from phase 1 to phase 4 clinical trials often involving cohorts of several thousand individuals). Targeted recruitment to and data transfer to/from these studies are essential to support.

It is important to note that prior to ENSAT-CANCER, adrenal tumor information was located in isolated national (and/or individual hospital) repositories of highly valuable data invisible to the vast majority of clinicians and researchers. These included the French COMETE (COrtico-MEdullo- Tumeurs Endocrines) (Plouin et al., 2008); the National Italian Study Group on Adrenal Tumors (NISGAT); the German Adrenal Network Improving Medical Education (GANIME) (Koschker et al., 2006)) and the United Kingdom Adrenal Cortical Tumour network (UK ACT)). The unification and harmonisation of these resources was key to the success of the ENSAT-CANCER disease registries (Sinnott et al. 2013; Stell & Sinnott 2012; Stell & Sinnott 2013).
The ENSAT-CANCER registries include extensive information covering a broad range of data related to patients with adrenal tumours (or suspected adrenal tumours). Figure 2 shows a subset of the information related to a single patient. Each patient has an associated set of data/forms collecting targeted information on surgical treatments that they have had; chemotherapy/radiotherapy treatments and associated dosages and time periods; the drugs that they have received through to the availability of biomaterials (blood, tissue, urine etc.). A key feature of ENSAT-CANCER is the longitudinal nature of the data sets. Many patients have extended history of treatments and/or follow up information. Tracking this and observing differences in the history and progress of disease sheds new insight into the treatment and management of patients – factoring in both the disease itself (and its stage); the treatment regimes that the patients are undergoing and the demographic and genetic uniqueness of the patient themselves.

Ideally all of the data collected in the registry would be of the highest possible quality. To ensure these data validation methods should be implemented as core elements of research database management systems, ideally with unified and standardized tool support. These tools must be mature, validated and interoperable. For many registries, important criterion can/should also be considered: imaging support, biomarker support, -omics analysis and bridges between experimental and clinical research data (Demotes-Mainard & Kubiak 2011; Ohmann & Kuchinke 2009).

Validation methods need to assess data quality, which will typically consist of data accuracy, data comparability, data timeliness and the overall registry completeness. This should be in accordance to published (accepted) frameworks with added focus in record eligibility and timeliness of the patient status (Arts et al. 2002; Baigent et al. 2008). For clinical data registries, the following criteria are some of the important issues to be able to assess the overall quality of data:

- the data is actually entered consistently for all patients, i.e. there is a measurable degree of completeness of the data;
- the data is entered correctly from a syntactic perspective, e.g. for fields where an integer was expected, an integer is actually given;
- the data is believable, e.g. the date of birth is not in the future and/or treatments are not given after the date of death;
- the data is semantically and scientifically believable, e.g. a pregnancy status of true is not possible for males;
- the data is regularly updated and tracks the clinical disease progression for the patients, i.e. for many conditions it is the longitudinal change in patients undergoing treatments that are essential to understand;
- the data incorporates the most recent information from the patients;
- the data is comparable with other records, e.g. when considered with other data sets in the registry for patients, outliers may be an indicator of errors that have been made with data entry;
- the data is actually useful!

Many of these issues can be enforced directly by software-driven data validation during data entry, e.g. dates in the past or in the future that cannot be possible. Indeed ensuring the syntactic validity of data entry is a measure of good software engineering practices. Many simple non-scientific data quality checks can be defined and enforced, however more detailed scientific data quality checks are challenging and typically require depths of domain knowledge. This is further exacerbated by the discrepancies and practices between sites: welfare systems and financial support vary between countries; treatment regimes can vary between sites; responses to treatment may vary between individuals - hence the need for registries such as ENSAT-CANCER. As a result, many of the data quality mechanisms identified above are more complex to define, deliver and enforce merely through software engineering checks. Furthermore, in complex registries like ENSAT-CANCER, demanding the first bullet point (data completeness) can be challenging from a range of criteria. Different centres collect different clinical information as part of their routine clinical care; some centres have a single data entry person whilst others may have multiple data entry personnel. The final bullet point (usefulness of the data) is arguably the most important point of all. There are many dimensions to usefulness that could be considered. One of the core ways that is considered in ENSAT-CANCER is the use of the data for recruitment of patients into targeted clinical trials. To discover new patterns and conclusions in biomedical research clinical trials depend explicitly on the quality of the data and empirically measuring its overall quality (Shah & Tenenbaum 2012). The ability to use data directly from the registry into clinical studies is thus highly desirable.

In this context work was undertaken to explore the quality of the data in the ENSAT-CANCER registry both with regard to the registry itself as well as its utility to meet the needs of targeted clinical trials. A major clinical trial that was used as the basis for this work was the EURINE-ACT study.

3 Case Study on Data Quality to Support the EURINE-ACT Clinical Trial

EURINE-ACT is a prospective longitudinal multi-centre diagnostic trial within the ENSAT-Consortium. The aim of this trial is to develop tools for accurate differentiation between malignant and benign tumours, to improve the early detection of recurrence of ACC and to better understand the biological processes that take place that lead to hormonal excess in adrenal tumours. The machine learning based approach to determine benign from malignant tumours from urine samples is described in (Arlt et al. 2011). Working is currently ongoing at the University of Melbourne to extend these algorithms to utilise both High Performance Computing facilities and Cloud-based resources. EURINE-ACT focuses on both patients with ACC and patients with benign lesions (NAPACA). To receive a prediction of the likelihood of an adrenal tumour being benign or malignant,
biochemical profiles of all relevant steroids are measured from 24-hour urine. The trial includes all patients with presence of an adrenal cortical mass and subsequently clusters them into three different categories (ACC, NAPACA, indeterminate). Figure 3 illustrates the data flow from the ENSAT-CANCER registry to the EURINE-ACT study.

A systematic centralized data monitoring was undertaken to review the data completeness (DC) and data accuracy (DA) of the EURINE-ACT study based on the data derived from the ENSAT-CANCER registries. That is, for those groups and sites contributing relevant clinical data to the ENSAT-CANCER registry, i.e. data on patients that meet the inclusion criteria for EURINE-ACT, what was the overall quality and accuracy of the data that was contributed? This work involved the visit of eight centres contributing to the ENSAT-CANCER registries across Europe.

### EURINE-ACT Study Flowchart

**EURINE-ACT Study Flowchart**

**UK-ACT**

- Clinical data entry into EURINE-ACT or NAPACA database
  - Clinical data entry into EURINE-ACT (If Patient Present, click “TO EURINE-ACT Code”)
- Collection of Biometric (If Biometric Form)
- 10 ml of a plain bottle 24-h urine (Include 48 hr collection volume in ml)
- 12 ml of a morning spot urine
- 1.5 ml serum
- 1.5 ml heparin plasma

**Adrenaloscopic Grooming (MO)**

- Clinical data entry into EURINE-ACT or NAPACA database
  - Clinical data entry into EURINE-ACT (If Patient Present, click “TO EURINE-ACT Code”)
- Collection of Biometric (If Biometric Form)
- 10 ml of a plain bottle 24-h urine (Include 48 hr collection volume in ml)
- 12 ml of a morning spot urine
- 1.5 ml serum
- 1.5 ml heparin plasma

**Figure 3: EURINE-ACT Data Flow**

EURINE-ACT had numerous explicit requirements on the data required in the ENSAT-CANCER registry. All of this data was defined and supported in the data model of the ENSAT-CANCER database, but as shown in Figure 4 it was often the case that these data sets were not given. Here ACTH is a hormone that is measured during a Dexamethasone Test (DST), a test that checks if an adenoma is hormone secreting (Nieman 2010). It is recommended to perform a DST in every patient including those patients with suspected ACCs (Terzolo et al. 2009). Figure 4 gives the example where a data entry person entered the value ‘Not Done’ for ACTH. An entered value can be defined as complete. But ACTH is an important criterion for external judgment about the behaviour of adrenal masses and can provide additional information regarding routine care. Figure 4 also shows whether imaging was undertaken. Similarly whilst completing the data requirement for CT Density ‘Not Done’ this is not clinically accurate information.

DC shows whether all necessary clinical information is entered into the registry. It does not measure/record whether the entered data is actually correct. Rather it gives information regarding the rigour of performance of data collection and the assiduousness of data entry of a researcher/research centre. When DC is low, critical gaps need to be identified in the data collection and data entry processes. This might result in the need for local training requirements for example. Where all sites have low DC for certain data items, this can be an indicator that this data item is not worth keeping in the registry or the sites don’t know that the item is mandatory for the trial. Often the reasons for incomplete information can only be found through a local investigation of data processes (on site monitoring). A rigorous evaluation of data completion was undertaken for those data items that were identified as essential for the EURINE-ACT study. The analysis was performed for both ACC and NAPACA records and the results shown in Figures 5 and 6.

### Figure 4: Example Items and Values of a DQS

<table>
<thead>
<tr>
<th>Item</th>
<th>Value</th>
<th>Completeness</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTH</td>
<td>Not Done</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>24h Urine Cortisol</td>
<td>[Select...]</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Imaging Form</td>
<td>&lt;exists&gt;</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Imaging Method</td>
<td>CT</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CT Density</td>
<td>Not Done</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Ki67</td>
<td>1%</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Weiss Score</td>
<td>Not calculable</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

**Score**

\[ \frac{5}{7} = 0.71 \]

\[ \frac{3}{7} = 0.43 \]

### Figure 5: Overall ACC DC

The DC of ACC records show a lack of pathology information. It is assumed that every patient without metastasis will undergo surgery and therefore will have information about the tumour tissue. A simply plausibility check was added, if the item ‘Distant metastasis’ is ‘No’, then the Weiss score (a cancer score based on several histopathological criteria), and the Ki67 score (a molecule detected in growing cells that can be used to gain an understanding of the rate at which the cells within a tumour are growing) must have entered values. In 193 patients without a spread of cancer cells, only 68% of pathology information was entered. Nearly all records showed no information about imaging.

Benign lesions (NAPACA tumours) are required to have imaging information. The entered values must therefore reflect this decision, so that an external reviewer can make a similar judgment based on the entered characteristics. In comparison to ACC records, the relevant EURINE-ACT data for NAPACA records show a better overall completeness level, however mandatory items about the detection of the tumour are still missing.
in every third case. Indeed the completeness check indicated that in nearly 30% of all patients no DST was performed nor it was not entered.

DC is an important aspect of data quality. Data quality is a combination of DC and DA. DA is a further criterion for the calculation of data quality for EURINE-ACT and it is important to determine whether the record meets the eligibility criteria for the study. The detailed data flow for patients EURINE-ACT is outlined in Figure 7.

Figure 7: Eligibility Conditions that apply on NAPACA Records

Figure 7 describes how for a NAPACA record, the biomaterial must be received at the EURINE-ACT central processing / analysis site (Birmingham) and stored in the biobank. If this condition is met, then it is required to check that imaging information is also given in the patient (registry) record - this can be a CT scan, MRI scan, PET scan or X-Ray (Blake et al. 2010). The next check is whether the patient underwent surgery. If this is the case, it is important to identify whether this was marked as an Adrenal Cortical Adenoma (ACA) or not. In the case of an ACA, it is mandatory to record information about the Weiss Score and the Ki67 Score. If all conditions are met, the record can be confirmed as eligible for EURINE-ACT. If one or more of these criteria are not met, it might be due to the data entry person not having this information at the time of data entry, or their lack of rigor or a variety of other possible causes.

Based on the analysis of the ENSAT-CANCER data, it was discovered that just around 50% of all entered NAPACA records could be confirmed as eligible for EURINE-ACT. However the majority of records did not include information about imaging characteristics. Additional problems also appeared in patients, who underwent surgery, e.g. there were only 15 of 28 records information about pathology given as shown in Figure 8.

Figure 8: NAPACA Eligibility Calculation

The analysis from the specific sites contributing to ENSAT-CANCER has illustrated a diversity of DA (completeness and meeting the criteria for eligibility). The DA score in Table 1 defines how many records from the ENSAT site records could be confirmed as actually eligible for NAPACA EURINE-ACT. The arithmetic mean of both values generates the data quality score (DQS). Centres with less than ten NAPACA records where not shown in this calculation, but considered in the total calculation.

Table 1: NAPACA Baseline DQS

Table 1 illustrates a diversity of data quality metrics across different sites (where the Centres on the left use country code, city codes and where required due to multiple centres in the same city, an integer). Thus PLWW2 is Poland Warsaw Centre 2; GRAT is Greece Athens etc.

The ability to strike the balance between overall completeness and data quality should be aimed at encouraging good behaviour of site staff entering clinical data into the ENSAT-CANCER registry. The more items demanded in the completeness calculation, the danger of...
potential lower data accuracy score. Alternatively mandating that all data is compulsory and has to be completely accurate is often difficult to enforce: many registries and studies can contain thousands or tens of thousands of data points per patient. A balance is thus needed. In this review of ENSAT centres a spectrum of data quality was discovered.

Through this work, the DQS was reported to every ENSAT centre at the end of every month. It was observed that between March 2014 and April 2014 the overall data completeness increased in the ENSAT-CANCER registry (Figure 9). It is questionable if this is caused by the regular data quality reports or by the site visits of monitors. The motivation of every trial member is still one of the most important criteria for an improvement of data completeness.

![Figure 9: NAPACA DC Improvements Mar-Apr 2014](image)

### 4 Conclusions and Further Work

The ENSAT-CANCER registry has been extremely successful in establishing a vibrant global research community working towards a better understanding of the underlying biology and management of patients with adrenal tumours. The registries that have been created (ACC, APA, NAPACA, Pheo/PGL) are used for over 20 major international clinical trials from phase 1 through to phase 4 studies often involving several thousand individuals with many thousands of data points per patient. Many of these trials and studies are linked directly with the registries and incorporate data feeds directly from the registry and in some cases, directly feed data back into the registries. The overall quality of the registry is essential. This work has shown that work still exists in further improvements in the data quality of the registries. Some of these improvements can be achieved through systematic monitoring and evaluation of data from sites and exploring processes to help them improve their local data entry processes.

Work is also ongoing to explore alternative software-based models for improving data quality. The first approach is to provide live information to data entry personnel. Knowing the completeness of the data that they are entering in real time is one avenue being pursued. Benchmarking their data completeness score against the average for the registry as a whole can help incentivize them into being more conscientious in their data entry. Figure 9 above illustrates that feedback to sites can indeed make a difference. The cost of continuous onsite monitoring is likely to be exorbitant in the longer term however.

A second model that is being considered is to have live feedback from the research community on the quality of the data that has been entered directly into the registry itself. Wikis, blogs and discussions related to lack of data or questions about the quality or accuracy of the data more generally are currently being explored. Key to these solutions however is to encourage improvements and not dissuade personnel from their assigned data entry tasks.

### 4.1 Acknowledgements

The ENSAT-CANCER project is funded through a grant from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 259735. The authors also gratefully acknowledge the ENSAT-CANCER and EURINE- ACT collaborators with special mention to Prof. Felix Beuschlein (ENSAT-CANCER principal investigator). We also acknowledge funding from the European Science Foundation that facilitated the data-monitoring site visits.

### 5 References


Khoury, M.J. et al. (2007): The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genetics in Medicine, 9*(10):665–674.


