Light Touch Identification of Cost/Risk in Complex Socio-Technical Systems

Iliada Eleftheriou, Suzanne M. Embury, Andrew Brass

Light Touch Identification of Cost/Risk in Complex Socio-Technical Systems

Iliada Eleftheriou, Suzanne M. Embury, and Andrew Brass

School of Computer Science, University of Manchester,
Oxford Road, Manchester, M13 9PL, UK
iliada.eleftheriou@manchester.ac.uk

Abstract. Information sharing within complex organisations is often source of considerable cost and risk. In previous work, we showed that points of highest IT cost/risk within an organisation are often located at the points where data moves from one context of use to another. We proposed a lightweight method of modelling the journeys data make within an organisation, and showed how to identify risky or costly boundaries. In this paper, we build on this previous work by evaluating the stability and completeness of the three core boundaries of our proposed method with staff from different clinical genomics hospital departments in the UK. Assessing our boundaries in the four new studies we found that although our core boundaries are stable in the new area of Clinical Genomics, domain-specific requirements of organisations can drive the need for additional boundaries. Finally, we discuss the feasibility of a general, low-cost process for identifying further boundaries of interest when applying the method to a new domain.

Key words: data movement boundaries, data journeys, information sharing, risk identification, cost identification

1 Introduction

Information systems in complex organisations are often sources of considerable cost and risk to their owners, as well as delivering business value. When resources must be stretched to deliver as much value as possible, we need to discover the places in the organisation where costs are higher than necessary and where value opportunities are being missed. However, by the same token, organisations can't afford to divert resources into expensive new developments or re-engineering activities. Any efforts to optimise current processes, or to create new processes, must be cheap and effective to carry out.

In previous work, we set out to design a lightweight method for identifying points of high cost and risk in organisations, in respect of the creation, management and use of data, that can be used for early stage decision making. That is, the method should not require a detailed model of the organisation or its processes to be produced before it can make predictions. Using 18 case studies of software failure and success in the UK National Health Service (NHS), we developed a method called *data journey modelling*, in which the broad movement

of data within and between organisations can be mapped [5]. Onto this model, we overlay information about key "boundaries": places in the journey where data moves to another context causing the use and interpretation of data to alter in a way that can endanger the portability of data and introduce high costs/risks to the organisation. For example, when data is moved out of the context of the department that created it, and is used by others, the perceived data quality often drops dramatically, since much of the metadata that allows it to be interpreted correctly is not gathered or transferred explicitly with the data.

We identified three core boundaries from the information in the case studies: boundaries between organisational sub-units; boundaries between human participants on different salary scales; and boundaries where the medium of representation for data changes (e.g., from digital to paper). We confirmed their usefulness in a retrospective evaluation of a cost-saving effort in an NHS radiography department [6]. All three boundaries proved quick and easy to apply, and were able to indicate the points which the human experts had selected for change in the cost-saving exercise, as well as a number of additional points where, the experts agreed, cost could have been saved. Having assessed the usefulness and accuracy of the boundaries, a number of questions arose. Would these boundaries be similarly effective in other domains? And are they a complete set, or are other boundaries needed to capture the major costs and risks?

In this paper, we describe our efforts to answer these questions, by working with staff from several different Clinical Genetics departments in the NHS. Clinical genomics involves a more complex patient pathway than we had observed in our previous work, and raises different data management challenges. We describe 3 new case studies in which our standard boundary set was able to predict points of cost/risk (Section 4). We also describe a further case study, in which we examine the completeness of our boundary set. We found types of cost that our core boundaries did not predict, and propose two new boundaries to cover the gap in this new domain (Section 5). Finally, we present a tentative method for the low-cost up-front identification of new boundaries, to ensure that any data journey modelling effort within new domains is equipped with the boundaries it needs to perform effective cost/risk identification (Section 6).

2 Related Work

Despite decades of research on risk management and cost estimation, still no lightweight decision making approach has been proposed for the early-stages of the development cycle. Existing techniques focus principally on creating detailed predictions based on substantial models of the planned development [2, 7, 8, 12, 13]. They support project managers throughout the development process itself, rather than giving a low-cost indicator for use in *early-stage* decision making.

Other modelling methods, like UML models, capture fine-grained flows, between low-level processing units within a system, making it hard to focus on higher-level aspects of the enterprise that can bring cost and risks, i.e. social factors [3, 11]. Other high-level approaches (e.g., Data Flow Modelling, Business

Process Modelling, and data provenance) can implicitly model data movements between a network of systems, they typically contain much more detail than is needed for our purposes [1]. They provide no abstraction as to which parts of the system should be captured for early-stage cost/risk identification and which can be safely ignored. Others, identifying and combining both technical and social aspects, require detailed organisational knowledge which is often hard, and time consuming to acquire [14, 15, 10]. Although they are powerful mechanisms to understand actors of an organisational infrastructure, they do not give us any specific means to identify likely costs.

In summary, we found no predictive approach that is sufficiently lightweight to be used as a decision-making aid in the very early stages of a development.

3 Background: Boundaries in Data Journey Modelling

To give the context for the work reported in this paper, we first give a brief overview of the data journey modelling approach. A data journey model is an abstract representation of the broad movements of data within and across collaborating organisations [5]. It is a socio-technical model in that it combines information about the movement of data between networks of people, as well as software systems of often different organisations. Since data journey models are intended for early-stage decision making (ideally before any implementation is initiated), they must be cheap to produce. This is achieved by keeping the focus on the broad information flows, rather than modelling the detailed processes of how data is captured, managed and used. Data journey models are intended to quickly show the major points of stress in organisational and cross-organisational data movement: the points where the meaning of data may inadvertently (and invisibly) change, where changes of media or governance requirements block the movement of data and where data may be lost in translation.

Figure 1 gives an example of a simple data journey model from a hypothetical NHS setting in which a general practitioner (GP) requests a blood test from a local pathology lab. As can be seen from the example, the data journey model shows the major containers (both electronic and physical) where data "rests" in between legs of its journey, from point of creation to point of use. Electronic containers are shown as cylinders and physical data containers as cuboids. Actors in various roles are also shown, indicating people or systems who interact with containers to create or consume information. The solid arrows indicate the legs of the journey where data moves between containers, while the dashed arrows indicate data creation and use by actors. Journey legs are labelled with the information type that moves along them.

The first step in the data journey modelling process is to create the data journey model. Our experience in working with a variety of NHS practitioners is that journey models can typically be created in 1–2 hours, and that clinical staff can create them as easily as those with an IT function. Although modelling the data journeys is only the first step in our technique, we found that it had value in itself, in helping health care practitioners (HCPs) to quickly gain an overview

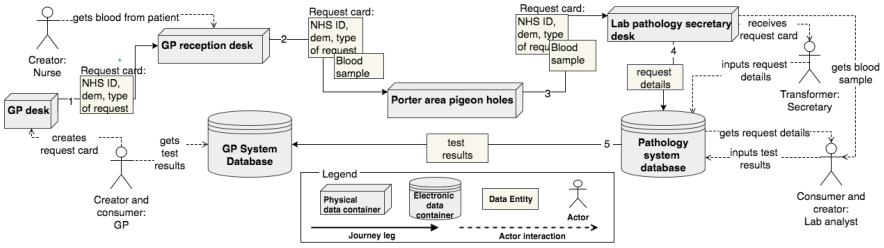

Fig. 1. Data journey model for a primary care blood test scenario

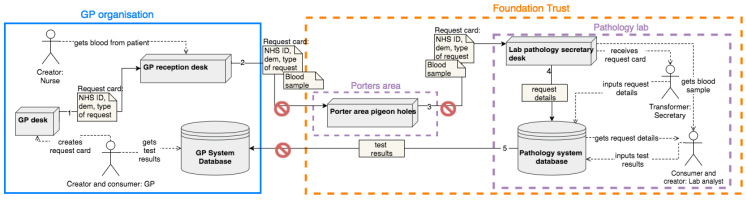

Fig. 2. Organisational boundary overlaid on the Pathology lab data journey model.

of how their organisation uses and shares data, when in their daily work they see only the detail of the parts of the journey they are directly involved with.

The second step is to add layers on top of the data journey model describing the key cost and risk factors likely to affect the successful movement of data. These layers group containers and actors that share similar characteristics relevant to the handling of data. The idea is that data movement within these groupings is likely to be low cost and low risk. However, when a journey leg takes data across a boundary between groups, then the risk of error, data loss, and additional data management costs increases. By overlaying several groupings/boundaries on a data journey model, a heat map of likely cost/risk points can be created. From our earlier work with 18 case studies of software failure and success, written by NHS staff, we extracted the following three boundaries:

- The 'Organisational' boundary, which groups containers and actors that are part of the same organisational unit (e.g. same department). Figure 2 shows the organisational boundaries overlaid on the pathology lab journey model.
- The 'Media' boundary, which groups containers that use the same media for information transfer and storage (e.g. paper files, x-rays, digital images).
- The 'Actors pay-scale' boundary, which groups actors who are at the same or similar salary level (e.g. clinician, secretary). This boundary is a low-cost proxy identifying groups of people of similar vocabularies and expertise, to identify points of information loss when data is shared across them.

The final stage of the model is to extract predictions of the places where operational costs might be high, or where the risk of significant costs occurring is high. This involves simply identifying the places where journey legs cross boundaries in the heat map, with legs that cross more boundaries being considered at

greater risk than legs which cross fewer. Stakeholders are then asked to consider the identified legs, and to identify whether a genuine cost (or risk of cost) is present. Since our method is intended for early stage decision making, no attempt is made to quantify the costs or risks predicted. The information needed for that is typically not available at this stage, and could be costly to acquire. Our aim is to quickly provide a broad brush picture of the likely costs and risks.

A previous study in an NHS radiology department demonstrated the ability of this method to correctly identify points of high cost within a real setting [6]. Despite the simplicity of our method, almost all the predicted optimisation points were found to be valid by hospital staff, with only 2 false negative results. However, this single study could not completely validate the data journey modelling technique, and a number of open questions remained:

- How 'stable' are the boundaries across different organisations and domains? Are these three boundaries capable of identifying points of cost/risk in other settings, or does each new setting require us to identify a new set of boundaries? If data journey modelling is to be truly low cost to apply, then we need to have a stable set of boundaries that can be used across many domains. We can't afford to have a lengthy boundary discovery process bolted onto our lightweight method.
- How 'complete' is this set of boundaries? Even if the three boundaries we have identified in our previous work are stable across domains, that leaves open the possibility that some important costs and risks are not being identified by our method, because they cross boundaries that we are not modelling.
- How expensive is it to determine possible important boundaries for new settings, before full data journey modelling has taken place? Even if we identify a set of boundaries that is stable and (largely) complete across domains, there is still the possibility that some particular area might have highly specific requirements that should be taken into account in the identification of costs and risks. Can we find a low-cost, up-front way to determine whether a setting has specific boundary requirements not covered by our standard boundary set?

We set out to obtain answers to these questions by working with groups of health care professionals in the Clinical Genomics area, a different domain than the one we had previously worked with. The processes we followed, and the results we obtained, are described in the following sections.

4 Stability of the Core Boundaries

In this section, we assess the stability of our method's three core boundaries in identifying places of high costs and risks in a new domains. To do so, we worked with health care professionals (HCPs) in the area of Clinical Genomics. Clinical genomics is a branch of medicine in which the genome of the patient is sequenced, and interpreted by a multi-skilled team of experts, in order to assist with diagnosis of hereditary conditions, inform treatment decisions, and to

determine the likelihood of conditions or symptoms appearing in the future [9]. The Clinical Genomics patient pathway is more complex than those we have worked with before, as it requires sharing of a rich variety of information between numerous actors, with very different skill sets, from different organisational units.

4.1 Study design

During the period covered by this paper, we were lucky enough to have access to three separate groups of HCPs, working in different hospital foundation trusts across the UK, and having quite different roles in the clinical genomics patient pathway (varying from managerial positions to specialist technicians). In working with each group, our research question was: are the core boundaries stable for this group? That is, if used alone, can our core boundaries identify some costs/risks deemed significant by the domain experts in domains different from the ones from which they were originally identified? We did not require that the core boundaries find *all* cost/risk points, only that significant ones could be identified.

We followed a method inspired by action research, but adapted to fit the circumstances of the opportunities we had to work with domain experts (some of which arose at short notice). In each case, we began by defining a procedure to follow, and took care only to document domain expert responses, and not to lead them to answers we might have wished they had given.

4.2 Clinical Genomics Study A: Phenotype Pathway

In the most extensive study of the three, we worked with staff in a Genomic Medicine department of a Foundation Trust hospital in Greater Manchester which had recently undertook a re-engineering phase to improve the efficiency of their processes. To evaluate the stability of our method's boundaries in this domain, we carried out a retrospective analysis of the data journeys needed to capture the patient's phenotype (the set of observable characteristics of the patient's genotype). We modelled the journeys of data *before* and *after* the information infrastructure redesign to assess the predictive power of our method's boundaries, using the following approach:

1. We conducted semi-structured interviews with a Bioinformatician in the Genomics department of the trust, to gather the necessary domain knowledge.
2. We modelled the *old* data journeys before any improvements were made by hospital staff and overlaid our core boundaries to identify the journey legs with likely high costs and risks.
3. Then, we modelled the data journeys of the *new* system after the hospital re-engineering team improvement efforts.
4. Finally, we compared the predicted costly journey legs of the old model with the staff's improvements to assess the stability of our method's boundaries in predicting places of high costs and risks. These improvements were at a sufficiently advanced stage to give some confidence that they were indeed cost-saving, though we had no way to confirm that in the scope of the study.

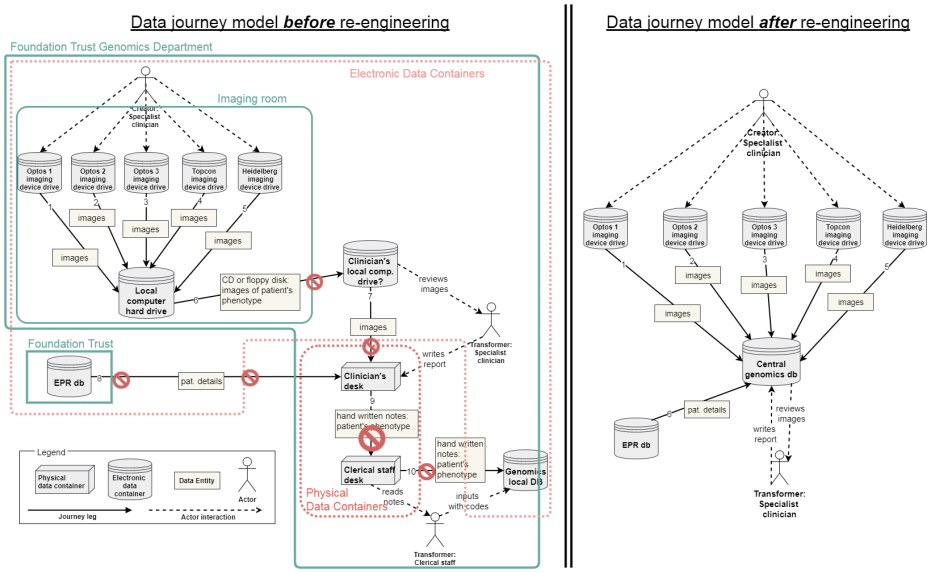

Fig. 3. Data journey model *before* and *after* the re-engineering process.

The journey models we produced from this exercise are shown in Figure 3. In the pathway we modelled, the phenotype of the patient is gathered by a specialist clinician from retinal images. The clinician examines these images and writes a report documenting the observed phenotype of the patient as hand written notes. Clerical staff then type up the reports. Before the re-engineering, several technical and social challenges caused considerate costs to the department. Diagnostic data was stored on a very old computer not connected to the network. To retrieve data, staff had to physically go to the room where the devices are stored and retrieve data using a USB drive, or sometimes even a floppy disk. Moreover, patient phenotype information was often needed by bioinformaticians and other actors working in the pathway, but no official sharing protocol existed, obstructing the dissemination of vital information. Another expensive issue was that phenotype data was not always coded (converted into standard medical codes) on input to the local computer system, making machine processing challenging.

With the help of the bioinformatician, we created the data journey model of the old system prior to the re-engineering phase, as shown in the left part of figure 3[1]. Once the bioinformatician approved the model, we overlaid the three core boundaries. In figure 3, organisational boundaries are denoted with a green colour solid line, the change of media boundaries with a red dotted line, and the change in pay-scale boundary with a large red warning sign. The journey

[1] As with many governmental institutions, aspects of this case study are confidential. Although the results presented were produced from the actual models, the illustrative models given in the paper are generalised, to show the typical journeys expected in a clinical genomics setting.

legs that cross the boundaries are legs 6, 7, 8, 9, and 10, indicated in the figure with a small red warning sign. The place of highest cost/risk as predicted by our method is journey leg no 8, since it crossed two of the three boundaries (organisational and media boundaries).

Next, we worked with the bioinformatician to model the current (improved) data journeys, as illustrated in the right part of figure 3. The new system replaced the old report-based process with a new computer system for storing patient phenotype electronically. Now the phenotype is stored in a central database for access by all members of the genomics team. Data captured by the retinal imaging devices are also uploaded to the central database.

Comparing the two models, we found that all but one of the predicted costly journey legs in the old model were removed in the current model. The one that remains (journey leg #8) in fact crosses two boundaries, and one of these (the media boundary) was removed by the re-engineering effort. The bioinformatician felt that cost savings were possible by removing the organisational boundary at this point, but that it would be challenging because this part of the journey was under the control of the wider Foundation Trust and therefore subject to hard-to-remove governance restrictions. Thus the three core boundaries were found to be stable and useful in this new domain, able to identify the same points of cost and risk that hospital staff identified and replaced.

4.3 Clinical Genomics Studies B and C

The next two studies were undertaken in conjunction with 4 Clinical Consultant Managers from different Genomics teams across the UK (unconnected with study A). The NHS staff members were participating in the doctoral level academic programme for Higher Specialist Scientist Training (HSST) at the University of Manchester. All were facing challenges in implementing new functionality in their trusts, to either integrate systems, migrate information to new servers, or expand existing functionality. To assess the stability of our method's core boundaries in their settings, we held a half-day workshop with them where we introduced the data journey modelling method, and asked them to use it to predict points of cost and risk in their departments. To do this, we:

1. Introduced data journey models as a tool to map the information infrastructure of an organisation and the data journeys happening within (without yet mentioning the predictive part of our method's boundaries).
2. Asked the clinical consultant managers to create models of the key data journeys happening in their departments.
3. Before mentioning the boundaries, we asked them to note on their models the journey legs they think are the most expensive (in terms of time, effort and resources) based on their experience.
4. Introduced the use of boundaries to predict places of high costs/risks, and asked them to apply the core boundaries to their data journey models.
5. Compared the places that clinical consultants assessed as most costly with the places our boundaries identified to assess their predictive power.

Working in pairs, the clinical consultants created data journey models corresponding to areas of concern in their respective departments. Each model was thus an amalgam of behaviours in two different Clinical Genomics departments in the UK. Each team focused (at their own choice) on slightly different parts of the patient pathway. Study B focused on the data journeys needed to collect information from several actors across the pathway for the bioinformatician to process. Study C focused on the data journeys needed to collect variant information from several external resources (e.g., Decipher and ClinVar).

Interestingly, all participants agreed that even this first step (of creating the data journey model) was valuable. Prior to the experience, they had a bioinformatics-centric picture of the work in their departments, since that was the focus of their everyday activities. The data journey model helped them to gain a different perspective on the processes and interactions taking place within their teams. In particular, the study B pair was surprised by the number of journey legs needed to collect required information for one key task in their pathway (exome assembly). The Study C pair focused on the external data sources that bioinformaticians rely on to feed their computational analysis pipeline. Before they started modelling, they planned to model the journeys data make from the external data sources to the bioinformatician's machine. However, while modelling these journey legs, they realised that the legs depend on a larger set of journey legs. Having created the data journey model, they recognised the value of other actors in the pipeline in getting accurate information to work with. The journey model brought home to all of them the complexity of the interactions between people and systems involved.

Having completed the data journey models, and before the boundaries had been added, we asked both pairs to identify the journey legs where most effort (cost) was expended, and where they were most concerned to see an efficiency gain. Having done this, we introduced the three core boundaries of our method and showed the participants how to layer the boundaries over the model and draw predictions from them. In both cases, the points of cost/risk identified by the boundaries differed from the problematic legs already identified by the teams. But upon consideration, both teams agreed that the predicted legs *were* sources of highest cost; they just hadn't been aware of them beforehand.

The case study B pair initially identified the bioinformatician's part of the model as the most costly, where the heavy data processing happens. The boundaries, however, identified a single point of failure in their current infrastructure — the majority of the journeys started and ended at a single system, maintained by one particular staff member. They had not realised the dependence of their processes on this one person before this exercise. Having identified it, they can look at re-engineering their processes and remove the single point of failure.

In case study C, the pair initially identified the legs moving data from the external sources to the bioinformatician's workplace as most costly, because of the need to pay data access fees. While the boundaries did predict these points of cost, they predicted a different leg as being of highest cost/risk: the point of information handover from the bioinformatician to the genetic counsellor,

which crossed all three boundaries. On consideration, both HCPs agreed that the point predicted by the boundaries was probably more costly, since it happens on a daily basis (the cost is accumulated daily), whereas the area they had been concentrating on happened more rarely.

Thus, in both cases (and in less than 2 hours, including training in data journey modelling) our 3 core boundaries were able to identify points of cost that the domain experts agreed with. In fact, in both cases, the predictions differed from the participants prior perceptions of the risks, giving them an insight into the needs of their departments that they did not have before. Taken together, studies A, B and C indicate that even this simple set of boundaries can have significant predictive power, at very low cost, in the Clinical Genomics domain.

5 Completeness of the Boundaries (Study D)

Case studies A, B and C allowed us to test the stability of our core boundaries in the Genomics domain. However, the nature of the case studies, and the small number of stakeholders involved, did not allow us to draw any conclusions about the completeness of this set of boundaries. This was the focus of our next case study, in which we set out to answer whether the core boundaries were able to identify most of the key points of cost and risk, or whether there are significant aspects of cost and risk that they cannot identify.

In this case study, we investigated the entire Clinical Genomics patient pathway, looking for significant costs/risks not captured by our 3 core boundaries. For this, we worked with the group of HSST staff from studies B and C, but also with another group of NHS staff attending the Clinical Bioinformatics Genomics Masters course at the University of Manchester. They came from various NHS Foundation Trusts in the Greater Manchester and Liverpool area, and worked in a range of positions within the Genomics patient pathway, such as Genetic Counsellor, Genome Technician, and Clinical Geneticist. To assess the completeness of the core 3 boundaries in identifying significant costs/risks, we:

1. Worked with clinical genomics staff to create a data journey model for the full patient pathway.
2. Observed the clinical bioinformatics module to collect socio-technical challenges across the entire patient pathway that NHS staff report facing in their everyday work and that cause significant costs for their departments.
3. Assessed each challenge to see if they were predicted by our core boundaries over-layed onto the patient pathway data journey model.

To look for costs/risks in the pathway, we asked the participants to discuss challenges they face in their everyday work. We collected not just technical challenges stemming from the data and technologies used, but also social and organisational challenges that all introduced some type of cost/effort to their everyday processes. The challenges identified from this process are summarised in table 1. We examined each challenge to determine whether costly/risky points

Challenges and potential costs/risks	Boundary
Data can be misunderstood by consumers when it is produced by teams with different backgrounds and expertise from the consumers.	Actor
Data that is used by one group of people but collected by another are often found to be inaccurate/incomplete. The consumer of the data in this case often experiences a decreased quality of the data.	Actor
Lack of communication between stakeholders and the development team.	Actor
Staff can be reluctant to share variants with other pipelines because of the governance frameworks in place.	?
Heterogeneity issues between external data sources. E.g. use of different IDs. Some sources may have older versions of the same data entity. Also, different sources represent the same data in different ways.	Org.
There are governance issues whenever data is transferred between networks of different organisations (university and hospital networks).	Org.
Information governance caused issues when integrating different parts of different pipelines (research and clinical).	?
Some data are 30–40 years old. Corrections to data over time cause duplicate versions of information (which are not explicitly marked at the source).	?
The sequencing machines produce very large volumes of data, causing storage and sharing problems.	?
Different bioinformaticians use different workflows to process the data, leading to potentially different results.	Actor
Needed information might not exist (i.e. never-seen-before variants). It can be hard to distinguish between non-existent data and data that is only absent from the source.	?
Information available in the literature is not always as accurate as claimed.	Org.
Clinical geneticists have only 10 minutes to comprehend findings produced by the bioinformaticians before making a potentially life-threatening decision.	Actor

Table 1. Challenges identified in the Clinical Genomics domain.

relating to it would be picked up on a data journey model using our 3 core boundaries. To do so, prior to this study, we had developed a comprehensive data journey model for the full clinical genomics pathway, with the help of the clinical genomics team at a nearby hospital. We looked at the points on this model where these challenges might materialise, to see if they coincided with the legs predicted to be cost/risk hot spots by our three standard boundaries. We present the boundary indicative of the cost described by each challenge to the second column of the table. Costly points not captured by any of the boundaries are noted with a question mark symbol.

From the table, we see that more than half the challenges (62%) can be identified by our core boundaries. The people-oriented challenges were all identified by our actor boundary, since it identifies points where data moves between members of staff with different kinds and degrees of specialist expertise. In our Clinical Genomics data journey model, this happens principally at the point of data handover from the clinical geneticists to the bioinformaticians, and *vice*

versa. These groups of people have very different specialisms, and do not always share the same understanding of the data.

The organisational boundary was able to indicate all the obstacles of sharing information with external sources. No change-of-media challenges were in evidence in the set of challenges we elicited from the domain experts in this case. However, some challenges were not identified by our standard boundaries. Specifically, those involving data volume and governance procedures were not captured by any of the standard boundaries. Both these factors limit the portability of data, and introduce additional costs to the patient pathway.

The data volume is a major obstacle to the movement of data between key actors in the pathway. Sequencing a patient's genome can produce millions of data files, resulting in large volumes of data (typically 10–100GB per patient). Given the complex nature of the pathway, with different actors typically working in different locations, and using their own dedicated software systems, large data volumes can be a real barrier to effective sharing and collaboration.

To validate the newly found obstacle of the volume of the data, we referred back to the case studies B and C. Examples of volume-related challenges were experienced by participants in both studies. One of the participants reported the need to establish a new journey to move information from an old data repository to a newly created one. However, no connection has been yet established to migrate information between the two repositories (there are no governance barriers, since the movement is within the same hospital). Apart from communication problems between the stakeholders and the development team, the major problem is the volume of data that needs to be shared. Since there is no direct network connection they currently have to copy each day's work (some 10 GB of data) to the new repository, through external hard disks, every night. Another participant reported the need for exome data to move between two geographically distant sites (two UK cities). However, exome data sets are typically around 5 GB, and attempts to transfer them cause the archive system which is used for transferring data to crash. Moreover, in another data movement example in a university hospital, data needs to be transferred from the external university network to the internal NHS network. Both machines are in the same room, but are on different networks. The volume of data to be transferred is large, and the network is slow. The participant reported that sometimes they have to plug the machine physically into the other network to transfer data.

The other new boundary retrieved from the challenges relates to information governance. In domains where information is highly private and confidential such as clinical genomics, information sharing must be tightly regulated and controlled. Governance protocols must be established to ensure that patient data is kept securely, and only used for agreed purposes. To complicate matters, governance protocols do not coincide with organisational boundaries. For example, within clinical genomics, there are two main protocols in use: a research oriented pipeline of processes and a clinical oriented pipeline. Each pipeline must follow the respective governance framework and guidelines. Audit information is captured along the pipeline based on the specific governance framework that

applies. If governance protocols conflict, or are not fit for purpose, serious delays and additional costs can affect data sharing efforts.

Thus, this study identified two additional boundaries needed to root out the key obstacles to data sharing in a clinical genomics setting: data volume and information governance. Although these boundaries arose out of highly domain specific situations, they are applicable across a broad range of domains. In the era of big data, clinical geneticists are not unique in having to work with much larger data sets than their IT infrastructure are designed for. Nor are information governance protocols limited to the handling of genetic data.

Since, it seems reasonable to assume that these boundaries will have wider applicability than just this one domain, the next step is to convert these high level concepts into actual "boundaries" that can be added to a data journey model. To be a useful boundary in our context, the information needed to decide which containers/actors are on which side of the boundary must be quick and cheap to acquire. To apply the volume boundary on the model, we group together containers storing data sets of similar size. A simple and quick way to categorise the volume of the data entities is by using the agile approach of 'tee shirt sizing', in which size is described by broad categories (small, medium, large) [4]. Then, boundaries show where data moves from containers handling large volume of data into containers set up to handle small volume, and vice versa. Similarly, to identify costs arising from the governance boundary, we group together containers and actors set to follow the research-oriented pipeline, and the clinical-oriented pipeline. Since containers and actors can work on both governance protocols, costs will arise when a journey leg moves data to a target container of a different protocol than those followed by the source.

Finally, identifying the new boundaries of data volume and information governance, suggests that domain-specific requirements of organisations might drive the need for additional boundaries during data journey modelling. The next section presents a new method to identify potential boundaries in other domains.

6 Identifying Boundaries in New Domains

Our work with the clinical genomics teams indicated that certain areas might have domain-specific costs and risks that may not be identified using our core set of boundaries. It seems likely that many new domains may share this characteristic: the generic boundaries can be applicable, but some domain-specific boundaries may also be needed. In this section, we discuss a method to identify additional boundaries to use when modelling data journeys in a new domain. It is an up-front approach that takes place before any data journey modelling is initiated, and low-cost, so as not to jeopardise the lightweight nature of the combined boundary-discover/data journey modelling technique. The process we propose for this is as follows:

1. "Grumble Analysis": hold a brainstorming session with stakeholders and domain experts to capture socio-technical challenges stemming from technical,

organisational, regulations, guidelines, and social aspects, that they face in their everyday work.
2. "Good/Bad Analysis": for every challenge identified, check whether an already established boundary matches it. If not, a new boundary may be waiting to be discovered. Ask the experts to suggest characteristics that differentiate participant organisational elements causing the problems described by the challenge, from those which do not. Primarily, we look for binary properties of the participants ("good" or "bad" characteristics) that are simple and cheap-to-acquire. For example, if data shared from external sources is often incomplete, we might distinguish sources based on their data admin response times. "Good" suppliers respond within 2 days, "bad" suppliers respond less promptly.
3. For each such characteristic, we look for cheap-to-acquire surrogates that can be used to form boundaries on the data journey model. An easy and quick way to do this is the tee-shirt agile approach mentioned above; categorise properties into simple broad classes, such as large, medium, small. Having categorised the properties of the participant organisational elements, we can group together those with similar size to form boundaries. Whenever data crosses a boundary, data moves from a source with a good property to a target of a different property that would impose costs to the journey.

Having identified a pool of potential boundaries, we then apply them on the data journey models we create within the new domain to check their usefulness in identifying costs and risks in the new domain. So, over time, we can build up a core set of boundaries to be used within that particular domain.

7 Conclusion

In this paper, we tested the three boundaries we had identified in previous work for stability across different settings, and for completeness. We applied the boundaries in three new case studies, in a different clinical domain than used in our previous work, and with staff in different roles in the patient pathway. We found that the boundaries performed consistently well across all three case studies; they were able to identify points of cost/risk that the staff agreed with, and (perhaps most importantly) were able to identify points of cost/risk that the staff themselves were not aware of before the modelling exercise. We also undertook a study of the entire clinical genomics patient pathway, looking for costs and risks that were not identifiable by our standard boundaries. We found that although many of the challenges were picked up by the standard boundaries, some were not suggesting that domain-specific organisational requirements can drive the need for additional boundaries. From this, we provide two new boundaries (data volume and data governance) to fill in the gap. We also present a low-cost technique for identifying new candidate boundaries in other domains, so that unique characteristics of the domain will not be missed during data journey modelling. In future work, we will check the applicability and power of our new boundaries,

as well as testing that our proposed method for discovering boundaries is indeed quick and effective to apply in other domains and other contexts.

References

1. R. S. Aguilar-Saven. Business process modelling: Review and framework. *International Journal of production economics*, 90(2):129–149, 2004.
2. B. Boehm, C. Abts, and S. Chulani. Software development cost estimation approaches - a survey. *Annals of software engineering*, 10(1-4):177–205, 2000.
3. Y. L. Chen et al. Data flow diagram. In *Modeling and Analysis of Enterprise and Information Systems*, pages 85–97. Springer, 2009.
4. C. G. Cobb. *The project manager's guide to mastering Agile: Principles and practices for an adaptive approach*. John Wiley & Sons, 2015.
5. I. Eleftheriou, S. Embury, and A. Brass. Data journey modelling: Predicting risk for IT developments. In *The Practice of Enterprise Modeling, 2016*. Springer, 2016.
6. I. Eleftheriou, S. Embury, R. Moden, P. Dobinson, and A. Brass. Data journeys: Identifying social and technical barriers to data movement in large, complex organisations. Technical report, School of Computer Science, UoM, 2016. Access link: www.datajourney.org/publications/tech_rep_data_journey.pdf.
7. M. Jørgensen and S. Grimstad. Software development effort estimation – demystifying and improving expert estimation. In *Simula Research Laboratory*, pages 381–403. Springer, 2010.
8. E. Mendes. Effort and risk prediction for healthcare software projects delivered on the web. In *Practitioner's Knowledge Representation*. Springer, 2014. p. 107–122.
9. D. W. Mount. Bioinformatics: sequence and genome analysis. *Journal of Bioinformatics*, 28, 2001.
10. T. Pardo, A. M. Cresswell, S. S. Dawes, et al. Modeling the social & technical processes of interorganizational information integration. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, page 8. IEEE, 2004.
11. Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *ACM Sigmod Record*, 34(3):31–36, 2005.
12. A. Trendowicz. *Software Cost Estimation, Benchmarking, and Risk Assessment: The Software Decision-Makers' Guide to Predictable Software Development*. Springer Science & Business Media, 2013.
13. A. Trendowicz and R. Jeffery. Principles of effort and cost estimation. In *Software project effort estimation*, pages 11–45. Springer, 2014.
14. E. S. Yu. Social modeling and i*. In *Conceptual Modeling: Foundations and Applications*, pages 99–121. Springer, 2009.
15. M. M. Yusof, J. Kuljis, A. Papazafeiropoulou, and L. K. Stergioulas. An evaluation framework for Health Information Systems: human, organization and technology-fit factors (HOT-fit). *International journal of medical informatics*, 77(6):386–398, 2008.