Aggregation and Reclassification - Assessment of GALEN Methods in the Domain of Thoracic Surgery

Mats Carlsson*, Jeremy Rogers†, Hans Åhlfeldt*
*Department of Medical Informatics, Linköping University, Sweden
†Medical Informatics Group, Department of Computer Science, University of Manchester, UK

This paper reports on the experiences from evaluation of GALEN methods for mapping of follow-up categories in the domain of thoracic surgery to an existing classification of surgical procedures. The mapping of the aggregated levels or groups of thoracic procedures presents a genuine problem in relation to strict hierarchical classifications, since the follow-up categories not necessarily fit in the pre-set structure of the classification. The paper reports on the experiences from modelling of the traditional classification as well as modelling of the follow-up categories, together with an analysis of results with a discussion of opportunities and potential problems and pitfalls when applying GALEN models and tools.

INTRODUCTION

How to classify is an age-old problem. Whenever a domain reaches a certain size there arises a need for categorizing, for dividing the domain into manageable pieces. There are many examples of this, especially within the natural sciences, of which medicine is but one example.

A common problem with traditional classifications used in medical subdomains such as diseases or medical interventions and procedures is their sheer size, making maintenance increasingly difficult. Another, and probably more serious, problem with these same classifications is their structure, namely the strict hierarchical classification. The limitations of hierarchical medical terminologies as abstracting systems is well documented in the literature[1]. Traditional hierarchical classification systems have been developed with a specific purpose in mind and are not well suited for reuse. Reuse becomes a necessity if clinical information is to be used not only for direct patient care, but also to support seamless care across healthcare organizational barriers, provide health statistics reporting, and facilitate follow-up and medical audit. Advanced terminological systems such as the GALEN terminology server[2], based on formal description of medical concepts and their relations with support for sanctioning mechanisms for composition of complex medical statements from atomic ones, promise sound solutions to the basic problems that arise from abstracting systems in the shape of traditional classifications.

The objective of this study was to evaluate GALEN methods and tools with respect to the problem of health statistics reporting in Sweden in the domain of thoracic surgery. Thoracic surgery, which is resource intensive, has increased in volume this last decade[3]. In order to evaluate these procedures a national database was created. By necessity this database has a more course grained classification than the regular classification of surgical procedures[4]. The interest is in mortality rates and complication rates for types of procedures rather than for the individual procedures themselves.

The paper is based on work done in the EU funded project GALEN-IN-USE (GIU). The goal for GIU has been to assist and facilitate collaborative work when constructing and maintaining classifications of surgical procedures[5]. Tools and methods from a previous EU-project, GALEN[6], has been used. The GALEN common reference model (CRM)[7] has also been used in this work.

The paper reports on an experiment to explore the use of GALEN tools to cross-map between a relatively detailed clinical reporting classification and a more abstract aggregation classification.

The classifications studied in this paper are: the Thoracic Surgery chapter (F) of the Nordic Classification of Surgical Procedures (NCSP), and a small national follow-up terminology for statistical reporting, also concerning thoracic surgery.

Since there are many demands to use classifications for diverse purposes there is a need for different groupings of the individual classification codes. Some are interested in what body parts are operated upon, while others focus on what kind of procedures are performed. There is therefore a need for being able to generate new aggregation levels from time to time. This need is compounded by the fact that new procedures are created and old ones are evolving over time.

MATERIAL AND METHOD

The core of GALEN

In GALEN the CRM is a central model of high level concepts that is used a start (or base) for the total medical modeling[7]. A set of top level concepts, such as *structures*, *substances* and *processes*, has been defined. These concepts have then been utilized as a core for building the ever growing model of medical knowledge represented in GRAIL (the GALEN Representation And Integration Language)[8]. The result is a compositional and generative model for medical terminology. The CRM also contains knowledge intended to restrict expressivity to that which is sensible to say, and thereby reject nonsense compositions.

An intermediate language

It was soon apparent that modeling directly in GRAIL[8] was too cumbersome for the majority of physicians[5]. They could do it, but only after a significant learning effort. Therefore an intermediate representation (ImR) was created[9,10]. This conceptual representation is easier to learn and use than GRAIL, being both less expressive and more relaxed. It can be (semi-)automatically expanded into GRAIL expressions. The ImR, and its GRAIL expansion algorithm, seeks to achieve a compromise between the requirement of computer systems for rigid formal representations and the human knowledge worker's preference for semi-formal, or completely informal, representations. The ImR is also suitable for validation work, both of modeling done in a particular center and between centers. It is a way to use the simplicity of what Rossi Mori[11] calls a 'second generation' system while still having the power of a 'third generation' system.

One advantage this representation has over GRAIL is that it is able to capture some concepts that the otherwise more powerful GRAIL to this date can not[9,10]. For example negation and the 'other' concept in the context of a certain classification.

Even though 'other' has a clear *conceptual* meaning in the GRAIL-model - it stands for a highly specialized and rather strange kind of negation, directly equivalent to the meaning of 'other' in a rubric. The problem is that 'other'-type rubrics need to behave in a magical way when trying to automatically derive a classification. In the current classification engine - i.e. in the current specification of GRAIL - there is no mechanism to achieve this magical functionality, so that although the meaning of 'other' is represented it does not behave as it should[8,9]. In GRAIL, children form a disjoint non-exhaustive partition of the level above[9] unlike a classification where the children must form an exhaustive partition of the level above[13]. Hence the latter captures the pragmatic information such as 'other', NOS (not otherwise specified) and NES (not elsewhere specified). This is not represented in GRAIL itself but in the way the specific classification is mapped to GRAIL.

Tools

The two most useful tools for the work described in this paper are the Surgical Procedure Editing Tool (SPET) and the Classification Manager (ClaM). Both are parts of the Classification Workbench (ClaW)[14].

The Nordic Classification of Surgical Procedures

As an example of a classification in a specific domain the Nordic surgical procedures (NCSP) has been chosen in this paper. In the early 1980's a study comparing surgical frequencies and activities was initiated by NOMESCO (Nordic Medico-Statistical Committee). This resulted in an abbreviated Nordic list of surgical procedures for Denmark, Finland, Norway and Sweden in 1989[4]. There are 20 chapters of which 15 are main chapters and the codes are structured in a strict hierarchy with four levels. There are about 7000 codes in total. NCSP is a traditional classification arranged in chapters and subchapters mainly according to organ systems (see figure 1).

Chapter F, which is the focus in this paper, contains the procedures dealing with the heart and major thoracic valves. There are 616 rubrics and the four levels are used. Most rubrics are about surgery on valves and blood vessels in, or connected to, the heart. There are also procedures for correcting arrhythmia and problems with impulse propagation.

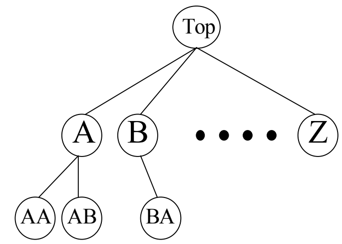

Figure 1. The structure of NCSP.

Aggregation for follow-up and statistical reporting

A national follow-up database related to thoracic surgery was set up 1992[3]. This database covers all clinics and units that perform heart surgery assisted by heart-lung machines. There are about 9000 patients per year. There are 12 units performing this kind of procedures in Sweden, and the reporting rate is 100%.

The surgeons have to report their deeds using eight fairly broad rubrics:

1. Valve surgery only
2. Coronary surgery only
3. Coronary surgery in other surgery
4. Congenital surgery
5. Transplanting surgery
6. Surgery for arrhythmia
7. Aorta-aneurysm surgery
8. Other heart surgery

To this date the years 1992-1995 have been reported. The patients are registered at the point of release from the care-unit. Half of the participating units reports by computer media. The report comes out annually, and has so far taken one year and three months to complete. The largest factor in the long completion time is that some units are late in reporting their data to the data base.

The data base is used for, and had as goal when created, to follow up and evaluate the result of heart surgery. The number of heart surgery procedures had been growing rapidly, and it was felt that some quality assurance was needed.

The follow-up codes could be categorized into three types. (I) the code covers one subchapter totally and only. (II) the code covers part of a subchapter, and finally (III) the code cover parts of more than one subchapter. Type I does not cause any problems when using NCSPs structure (figure 1). It is type III that is problematic and where a concept system as such CRM could be of help. Type II is not unproblematic but positions itself between I and III and can benefit from a concept system.

RESULTS

The reclassified classification

The process of modeling into the ImR is fairly straight forward. The English and the Swedish rubrics were entered. Then the paraphrase was decided upon and the modeling done from that. The paraphrase is an attempt to formulate what the rubrics really means. The SPET can generate a paraphrase in natural language from the for checking if the modeling is sensible.

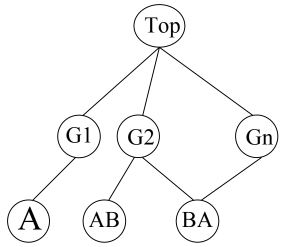

Figure 2. The CRM structure.

To reclassify, the ClaW is used. The hierarchical structure of NCSP was presented to the system (through the notion of chapters, subchapters and codes), where after all rubrics of chapter F were modeled and compiled into the GRAIL model. After these steps, analysis of NCSP could be performed, allowing the GRAIL classification engine to reclassify the NCSP codes according to their position within the CRM. The new follow-up rubrics could then be modeled and compiled into the GRAIL model with the assumption that they would subsume the expected number of NCSP codes according to the intention of the follow-up categories.

Most of the F-chapter was unproblematic, but the follow-up-rubrics did cause some problems. The 'only' statement was possible to represent in the ImR (see figure 3), but this has no meaning in GRAIL and would be expanded to figure 4. The 'only' just results in a specification of figure 5.

```
RUBRIC "Endast klafflkirurgi"
PARAPHRASE "surgical deed valve of heart, surgical deed"
ENGLISH_RUBRIC "Valve surgery only"
SOURCE "NCSP" CODE "FU1"
COMMENT "Quality assurance"
MAIN surgical deed
    ACTS_ON valve
        IS_PART_OF heart
EXCLUDES_CONTEXT OTHER surgical deed
```

Figure 3. ImR.

```
Process which
 involves Valve which
  isSolidRegionOf Heart
 isCharacterisedBy NON Process which
  hasDissectionCodingArtefact CodingOtherness
—extrinsic—
hasDissectionRubric extrinsical 'NCSP FU1 Endast klafflkirurgi'
```

Figure 4. The corresponding GRAIL-code.

```
Process which
 involves Valve which
  isSolidRegionOf Heart
```

Figure 5. GRAIL-code.

The 'isCharacterisedBy NON Process which hasDissection-CodingArtefact CodingOtherness ' statement does not mean anything in GRAIL. Just as 'other' it has yet to be defined. Also, negation 'NON' is not implemented. The 'only' part of the follow-up codes has been ignored thus making follow-up code two and three equivalent, for the purpose of this study.

The new aggregated levels

The follow-up rubrics (Fu-rubric in table 1) were modeled and compiled into the GRAIL model. This was done after the F-chapter had been compiled. Following this a reclassification was done. The result is shown in table 1. A thoracic surgeon was asked to give an interpretation of how the follow-up codes are used locally by grouping the subchapters in NCSP, on the three character level, under the eight follow-up rubrics. The 'miss' and 'extra' columns in table 1 are related to the local usage (LU).

Fu-rubric	Catch	Miss	Extra	LU	Type
1	100	4	15	78	III
2	59	4	9	55	I
4	50	125	0	175	III
5	14	2	0	16	I
6	13	15	1	27	II
7	6	46	5	47	I
8	-	-	-	-	III

Table 1. Result of the comparison

DISCUSSION

The follow-up rubrics did get varying results. Number one did well, with only four missed NCSP-codes. Furthermore three of those was not in the GRAIL-model at all due to the missing mapping for the ImR link WO ('WITH_ OPTIONALLY' it lacked a mapping into GRAIL) . The 15 extra where in accordance with the GRAIL-model since they all concerned surgical deeds on heart valves, so they could be said to be correct also from a medical perspective. The one doubtful code is FCE00 which concerns not a valve itself but a valve cavity.

Follow-up rubric two also managed well. Two of the missed codes were caused by the WO-link. The other two were procedures peripherally connected to coronary arteries: circumflex of the coronary artery branch, and artery connected to the left anterior descending coronary artery. The extras were all, except one, procedures for making a bypass to the coronary artery, and consequently arguably should be in this group. The last one was related to the coronary sinus.

For follow-up rubric three it became obvious that there is also the problem of interpreting the original follow-up rubrics. The semantics are not always clear. This is reflected by the fact that our surgeon did not group anything under this rubric.

Nine of the misses for follow-up rubric number four were caused by the WO-link cause, but the rest were caused by other things. Phenomenons like stenosis and chronic lesion, as also planar defect was not subsumed by the rubric, probably because they do not have to be congenital. But also congenital lesion was missed.

The two missed by rubric five was due to the modeling in ImR. They where modeled as surgery that occurs *during* transplanting, not as transplanting as such, thus not being subsumed by the follow-up rubric.

Rubric six had only one extra code as compared to the surgeons model. But this is due only to the fact that this code was explicitly excepted from the group that the rubric should catch. From the GRAIL-model point of view the subsumation was correct. The aggregation missed cardiac dysrythmia and atrail fibrillation plus 'EctopicCardiacDepolarisingFocus' and 'DisorderOfMyocardialConduction'.

The poor result of rubric seven is due to how the NCSP codes concerning aneurysms have been modeled in SPET. Often the descriptor 'aneurysm' has not been used, thus causing the follow-up rubric to fail to subsume. This is a case of semantic omission when the NCSP-codes either state or imply by their position in the classification that they concern aneurysms, but the modeler has not taken this into account. Of the extra ones only one is correct, the other are aneurysms in the heart. They were subsumed due to the necessity to construct a more general follow-up rubric in order to subsume anything at all.

The fact that rubric eight did not catch anything is due to the handling of 'other' in GRAIL. This was known and was thus ignored for the scope of this study.

Some NCSP rubrics were subsumed by more than one follow-up rubric. This follows logically from how the GRAIL model is built and was also the case in the local usage of the follow-up codes.

The main difference comes from the fact that NCSP and CRM have different structures. The classification is structured by organs and organ systems, mostly. But there are also chapters especially for endoscopic procedures and minor surgery[4]. The CRM is structured according to anatomic structure[7]. Therefore it is natural that some differences occurs (compare figure 1 and 2).

One problem in the original classification (NCSP) besides the sometimes unintuitive structure is the granularity. It does not always correspond with what the physicians want to express. Sometimes the granularity is insufficient and sometimes it is too detailed. On the other hand, the CRM can be hard to understand without training. On a superficial level it is quite easy to understand, but on a deeper level, the subtleties can be very hard to grasp.

On the whole it can be said that what was captured by the follow-up rubrics seems reasonable, but there was a significant set missed. It seems like the sensitivity of how rubrics were modeled were great. Also, since many factors are influenced by how the CRM is modeled it can be hard to know which modeling strategy will yield what result. A certain amount of trial an error is involved, if the modeler does not have insight into how the CRM is modeled.

One interesting thing is the modeling of heart valve in follow-up rubric one. Three different modelings where used, with varying results. At a superficial level these three modelings may seem analogous:

1 valve HAS_LOCATION heart
2 valve IS_PART_OF heart
3 heartvalve

becomes in GRAIL:

1 Valve which <involves Heart>
2 Valve which <isSolidRegionOf Heart>
3 MajorHeartValve

However, when looked at closely there are some semantic differences:

valve IS_PART_OF heart

...would subsume all of the major heart valves, plus the in utero one, and all mechanical or biological implant valves. The common thread is that they are all valves, and they are all in some way 'part of' the heart.

valve HAS_LOCATION heart

...would not subsume any of the above, but might subsume the notion of an artificial venous valve originally sited somewhere else in the body but which has become dislodged, floated up the venous system until it enters the heart and becomes stuck there. In such a situation, such a valve would be located in, but not part of, the heart.

Some NCSP rubrics having the (apparent) same GRAIL code end up as separate nodes in the GRAIL model. If looked upon in the full GRAIL-notation it becomes clear they have slightly different contexts. I.e. 'hasContextOfCodeNamed 'FGW96'' and 'hasContextOfCodeNamed 'FGW''. This might seem irrelevant at first glance, but the relevance becomes apparent when the place of the rubrics in the classification is taken into account. The information might not be useful in GRAIL itself, but can be used by an application utilizing a terminology server build on GRAIL. So even if the 'context' attribute does not change anything in how the GRAIL code is handled by the terminology server, the information is useful for a classification application.

One way to use the GALEN tools is to devise new follow-up rubrics, starting from the structure in the CRM. Finding alternative follow-up rubrics can be a fairly complicated task, since the knowledge embedded in the CRM is not always readily accessible without deeper insight in how the model is built.

The ImR works well, specially in conjunction with the dedicated tool. The problems that arise might have more to do with the SPET than the ImR per se. Descriptors may only be placed at one position in the descriptor hierarchy, which sometimes leads to problems of expressional power. It becomes cumbersome to say what you want to say in some cases. Most of the time this is a case of keeping to a certain style in order for the classification codes to be compiled into the correct place in the GRAIL-model. But there are also examples of when the ImR becomes unnecessarily hard to do. The benefits of structure and mapping to GRAIL outweighs this. In the few cases where it is relevant the mapping from ImR to GRAIL can be made by hand, by an individual so trained.

The exercise of finding which style is most beneficial to model in is an iterative process, and will be reached by discussions between coding centers. An example of this is the bypass procedure. First it was modeled by some as a surgical deed ('bypassing') that acted on some coronary blood vessel. But this caused problems with the procedures for removing bypasses. So it was agreed that the deed is 'creating' a bypass structure, which then can be removed.

There seems to be a general agreement that formal systems are a good thing, but there is also an assumption that those formal systems are going to have to represent faithfully all the existing systems. The fact is that the existing systems can not be represented faithfully because they are informal, inconsistent or use logical constructs like negation that are basically extremely difficult to compute in the worst case. 'other' is an obvious example of this. I.e. in Read 3 such things as NOS are marked 'optional' and can be filtered out. This since they are deemed not clinically useful by the specialty working groups[15]. These concepts usually are remnants from earlier versions and especially residuals from formal classifications.

There is a movement towards removing NOS, NES, and 'other' from classifications[15,16]. But all do not agree, some argue that the classifications are complementary to concept systems and due to their different purpose classifications need NOS, NES and 'other'[13]. This is because they are used statistically and to answer specific questions.

CONCLUSION

In regard to the aggregation of new follow-up codes, or validating old ones, the GALEN tools are helpful. This especially for type III, for which there are no support in classical classifications, like NCSP. Also when devising new follow-up codes the kind of facilities the GALEN tools provide are of help.

The reasons for NCSP codes to end up in the wrong place in the GRAIL-model and the follow-up rubrics not subsuming what could be expected are several.

Firstly, modelers make mistakes when modeling in the ImR, i.e. by semantic omission as for follow-up rubric seven. There might also be a certain amount of inconsistency in their modeling.

Secondly, negation in different forms, such as 'other', 'only' and 'not' causes problems for formal systems, but are useful for statistical classifications.

Thirdly, the mapping from ImR to GRAIL is opaque for the modelers, making it hard to predict the behavior of the modeled codes.

Lastly, the CRM is not easy to grasp in all its glory and complexity. This leading to misunderstandings and misinterpretations from the ImR-modelers side.

In spite of these problems our study indicates the potential power of the GALEN tools in that they provide modelers with methods for handling of some of the inherent problems with strict traditional classifications; lack of multiple views and flexibility in generation of new aggregation levels.

Acknowledgements

With thanks to all in the GALEN-IN-USE consortium. GALEN-IN-USE is founded as part of Framework IV of the Healthcare Telematics research program

References

1. Cimino JJ. Coding systems in health care. IMIA Yearbook of Medical Informatics 1995, 71-85.
2. Rector AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, Claassen WM. A Terminology Server for medical language and medical information systems. Methods Information Medicine. 34(1-2):147-57, 1995.
3. Socialstyrelsen. Svenska hjärtkirurgiregistret. Medicinsk faktadatabas MARS, 1997: http://www.sos.se/mars/kva006/k06.htm
4. NOMESCO. Classifications of Surgical Procedures. Nordic Medico-Statistical Committee 1996.
5. Galeazzi E, Rossi Mori A, Consorti F, Errera A, Merialdo P. A cooperative methodology to build conceptual models in medicine. Medical Informatics Europe '97, IOS Press, 1997: 280-284
6. Rector AL, Nowlan WA, the GALEN Consortium. The GALEN ProjecComp. Meth. and Prog. in Biomedicine, 1994: 75-78.
7. Rector A, Rogers J, Pole P. The Galen High Level Ontology. Medical Informatics Europe '96, Ed: Brender J, et al. IOS Press, 1996: 174-178.
8. Rector A, Bechhofer S, Goble C, Horrocks I, Nowlan A, Solomon D. The GRAIL concept modeling language for medical terminology. Artificial Intelligence in Medicine 1997; 9: 139-171.
9. Rector A. Thesauri and Formal Classifications: Terminologies for People and Machines. Meth Inform Med 1998; 37: 501-509.
10. Rogers J, Rector A. Terminological Systems: Bridging the Generation Gap. Annual Fall Symposium of American Medical Informatics Association. Nashville TN, Hanley & Belfus Inc. Philadelphia PA,1997: 610-614.
11. Rossi Mori A, Consorti F, Galeazzi E. Standards to support development of terminological systems for healthcara telematics. (proocedings of IMIA Working Group 6 meeing, Jacksonville, Florida).
12. Rogers J, Solomon W, Rector A, Pole P, Zanstra P, van der Haring E. Rubrics to Dissections to GRAIL to Classifications. Medical Informatics Europe '97, IOS Press, 1997: 241-245
13. Ingenerf J, Giere W. Concept-oriented Standardization and Statistics-oriented Classification: Continuing the Classification versus Nomenclature Controversy. Meth Inform Med 1998; 37: 527-539.
14. van der Haring E, Zanstra P, Claassen W. Classification Manager (ClaM IV). GALEN-IN-USE Deliverable, Prototype, Internal document, 1997: Deliverable 7.2
15. Campbell J, Carpenter P, Sneiderman C, Cohn S, Chute C, Warren JPhase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity. JAMIA 1997; 4: 238-251.
16. Robinson D, Schulz E, Brown P, Price C. Updating the Read Codes: User-interactive Maintenance of a Dynamic Clinical Vocabulary. JAMIA 1997; 6: 465-472.