Biomedical Informatics: We Are What We Publish

P. L. Elkin¹; S. H. Brown²,³; G. Wright⁴

¹Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA;
²Department of Veterans Affairs, Washington, DC, USA;
³Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA;
⁴Faculty of Health Sciences, Walter Sisulu University, Mthatha, South Africa

Keywords
Medical informatics, bioinformatics, publication record, expert consensus, ontology

Summary
Introduction: This article is part of a For-Discussion-Section of Methods of Information in Medicine on “Biomedical Informatics: We are what we publish”. It is introduced by an editorial and followed by a commentary paper with invited comments. In subsequent issues the discussion may continue through letters to the editor.

Objective: Informatics experts have attempted to define the field via consensus projects which has led to consensus statements by both AMIA and by IMIA. We add to the output of this process the results of a study of the Pubmed publications with abstracts from the field of Biomedical Informatics.

Methods: We took the terms from the AMIA consensus document and the terms from the IMIA definitions of the field of Biomedical Informatics and combined them through human review to create the Health Informatics Ontology. We built a terminology server using the Intelligent Natural Language Processor (INLP). Then we downloaded the entire set of articles in Medline identified by searching the literature by “Medical Informatics” OR “Bioinformatics”. The articles were parsed by the joint AMIA/IMIA terminology and then again using SNOMED CT and for the Bioinformatics they were also parsed using HGNC Ontology. We identified 153,580 articles using “Medical Informatics” and 20,573 articles using “Bioinformatics”. This resulted in 168,298 unique articles and an overlap of 5,855 articles. Of these 62,244 articles (37%) had titles and abstracts that contained at least one concept from the Health Informatics Ontology. SNOMED CT indexing showed that the field interacts with most all clinical fields of medicine.

Conclusions: Further defining the field by what we publish can add value to the consensus driven processes that have been the mainstay of the efforts to date. Next steps should be to extract terms from the literature that are uncovered and create class hierarchies and relationships for this content. We should also examine the high occurring of MeSH terms as markers to define Biomedical Informatics. Greater understanding of the Biomedical Informatics Literature has the potential to lead to improved self-awareness for our field.

1. Introduction

Biomedical Informatics is an inherently complex and interdisciplinary field. The interdisciplinary nature of the field accounts for much of the richness of the field. Those of us who have practiced the specialty have come to cherish this diversity. The same diversity makes the field more challenging to define. Each clinical and scientific subset of Informatics professionals has a unique and valid perspective of the specialty.

In spite of the challenging nature of this endeavor, AMIA and IMIA have each individually taken steps to define the field. Reinhold Haux et al. defined undergraduate education in Medical Informatics in a landmark article [1]. AMIA published their core competencies (Figure 1) after a set of meetings in 2003 which worked to formally define concepts in the field of Biomedical Informatics.

Arie Hasman recognized that Informatics curricula must meet the needs of students aimed at other than pure research careers [2]. In 2002, Musen published that construing work in medical informatics in terms of actions involving ontologies and problem-solving methods may move us closer to a theoretical basis for our field [3]. In 2004, Graham Wright brought together a group of international Informatics experts to develop systematic definitions for the field of Medical and Health Informatics over a four year period, which included an analysis of keywords used in publications (Figure 2 and Figure 3), however they excluded bioinformatics from this work [4]. In 2009, Schuemie and colleagues found that the Medical Informatics literature could be divided into three subdomains: 1) the organization, application, and evaluation of health information systems, 2) medical
knowledge representation, and 3) signal and data analysis [5]. In 2011, Reed Gardner published the definitions of the specialty of Clinical Informatics [6] in a consensus statement and in 2012 Kulikowski, Shortliffe et al. published AMIA’s whitepaper defining the field of Biomedical Informatics [7].

Top down approaches to defining Biomedical Informatics have the advantage that they are systematic and involve a broad set of stakeholders. The downside of this expert consensus based approach is that it is not data driven. In this manuscript, we attempt to understand if the published Biomedical Informatics literature can help us to better understand the breadth of the field of Biomedical Informatics.

With the advent of the new sub-specialty of Clinical Informatics [8] approved by the American Board of Medical Specialties (ABMS) and sponsored by the American Board of Preventative Medicine and the

Figure 1 AMIA Consensus Panel Core Competencies for the field of Biomedical Informatics shown in the TMU

Methods Inf Med 6/2013 © Schattauer 2013

539

P. L. Elkin et al.: Biomedical Informatics: We Are What We Publish
American Board of Pathology, we have seen an increased interest on the part of leaders in our specialty to further define Biomedical Informatics and to specify its core curriculum. The first board examination was administered on October 10, 2013. ABMS approved fellowships which will be available to any physician that holds a current board certification, which will soon be available to U.S. trainees.

It is expected that physicians trained as Clinical Informaticians will in some cases serve as the CMIO or work in the CMIO’s office for healthcare organizations. These groups and individuals provide strategic guidance to the leadership of health care organizations regarding current and future developments in Health IT and more broadly in Biomedical Informatics.

The iNLP software used in this study has been previously shown to have a very high performance with a recall (sensitivity) of mappings of 99.7%, a precision (positive predictive value) of 99.8% and a specificity of 97.9% [9].

2. Methods

The AMIA and IMIA terminologies were analyzed and merged. The AMIA terminology had 287 unique concepts and 315 terms. The IMIA terminology had 333 concepts and 348 terms (Figure 4). They were merged by hand using the Terminology Merge Utility (TMU) an open source utility created in the Laboratory of Biomedical Informatics directed by Dr. Elkin. Dr. Elkin performed the merge finding conceptual equivalence where possible.
(187 concepts) and creating new hierarchies when the content was unique to one of the contributing classifications. When merged this resulted in a Health Informatics Ontology (HIO) containing 433 concepts and 462 terms (Figure 5). A graphical view of the HIO is shown in Figure 6.

The resultant Health Informatics Ontology was built into a terminology server using the intelligent natural language processor (iNLP) [10]. This system was developed at Mount Sinai School of Medicine and is built upon the .Net architecture from Microsoft and is written in C#. The system is terminology and language independent and runs in a four-tier architecture. The system has been used successfully to identify post-operative complications from clinical notes [11] and as a method to improve the United States’s national biosurveillance strategy [12].

In this project, we used the iNLP to build a terminology server capable of parsing free text with the merged AMIA-IMIA Health Informatics Ontology (HIO). Searches of the medical literature were performed using Pubmed from the National Library of Medicine in February of 2006. The search criteria employed was “Medical Informatics OR Bioinformatics” exactly. Then each of the 168,298 titles and 121,561 abstracts were parsed with the HIO using the iNLP terminology server. The results were stored in a SQLServer database.

A random sample of 27,000 articles with abstracts from the literature was parsed with the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) to determine the breadth of clinical coverage attributable to the Informatics literature. We chose this number as it was approximately the number of records which could be parsed in a three day (24 hr) period. Figure 7 has an example of a conclusion from an informatics article [12] parsed with SNOMED CT by the iNLP server.

The Bioinformatics articles were also parsed with the Human Gene Nomenclature Committee Ontology (HGNC Ontology) to determine the breadth of human gene coverage attributable to the Informatics literature and how well HGNC Ontology represented the content of bioinformatics articles.

Descriptive statistics were applied to the resulting dataset of codified Informatics literature.

### 3. Results

The query “Medical Informatics” in Pubmed returned 153,580 articles and the

Figure 3  Example as to how the IMIA Classification of Biomedical Informatics is organized

Figure 4  Terminology Merge Utility showing the highest level categorizations for both the IMIA and AMIA Representations
query “Bioinformatics” returned 20,573 articles, or in total 174,153 articles. The query “Medical Informatics OR Bioinformatics” in Pubmed returned 168,298 unique records. Therefore the overlap between the two queries was 5,855 articles were indexed by both concepts or contained both keywords.

Of the 168,298 unique articles, 121,561 had abstracts and 14,645 were review articles. Of the 153,580 articles indexed as “Medical Informatics”, 12,279 were review articles and of the 20,573 articles indexed as “Bioinformatics”, 3,091 were review articles.

The iNLP HIO found at least one concept to be associated with 62,244 of the 168,298 articles (37%). Conversely 63% of the articles titles and abstracts were uncovered by the HIO. When indexed on average an article contained 2.32 HIO codes. Only 251 of the 433 concepts from the HIO were identified in the Informatics literature. The top terms identified in the informatics literature using the HIO terminology server are shown in Table 2.

Of the 20,573 Bioinformatics articles, 14,427 had HGNC Ontology codes. Of the 26,953 human genes only 3,275 were identified in this corpus of Bioinformatics literature.

When we combine the HIO and HGNC Ontology we find coverage of 76,671 (45.6%) of the 168,298 unique articles.

To evaluate the overlap of Biomedical Informatics with clinical medical subspecialties we parsed 27,000 articles with abstracts and the most frequent areas by body site identified are listed in Table 3. Overall there were 37,141 occurrences of SNOMED CT clinical concepts in these 27,000 titles and abstracts. Body site areas without representation in the Biomedical Informatics literature are listed in Table 4.

4. Discussion

The Biomedical Informatics literature has important content to contribute to our understanding of the breadth and depth of the field of Biomedical Informatics. The top down approach to modeling our field is useful but incomplete and could be advanced by efforts to better understand our literature. In this study of a large corpus of titles and abstracts, only 37% of articles titles and abstracts were identified using the Health Informatics Ontology.

Medical Informatics accounted for many fold more articles than Bioinformatics. SNOMED CT indexing showed that the field of Biomedical Informatics has broad involvement with most clinical specialties. This integration with the larger medical scientific community shows a maturing of the field as a health and biomedical scientific discipline.
The IMIA and AMIA consensus driven expert designed classifications are both authoritative. We are not suggesting that these necessarily be combined but we do believe that it is a testament to the hard work and diligence of these two communities that there was enough consistency that we were able to merge the two classifications.

As our field matures we are increasingly challenged to define our field both for our own internal uses (e.g. curriculum development) and to the biomedical community at large. It is our hope that our literature can help us to be more complete in our description of the breadth and depth of the content of our specialty.

### 5. Limitations

A limitation of these results is that the Biomedical Informatics literature is ever increasing and changing and therefore these results represent only a snapshot of the literature in 2006. There are alternate literature searches that might have been employed. There can be false positive articles retrieved in the results from Pubmed queries, so our numbers are likely not generalizable. We did not index the full text of articles but only the titles and abstracts. Abstracts are more often structured and more and more often convey the more important aspects of a paper’s substance. The basic conclusion that articles that are indexed as being either with “Medical Informatics” or “Bioinformatics” are not indexable by one of the terms that we have used to define our field is likely correct. We have not yet taken the next step to see what

### Table 1  AMIA and IMIA terminologies and the merged Health Informatics Ontology

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Concepts</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMIA</td>
<td>333</td>
<td>348</td>
</tr>
<tr>
<td>AMIA</td>
<td>287</td>
<td>315</td>
</tr>
<tr>
<td>Merged IMIA-AMI Health Informatics Ontology (HIO)</td>
<td>433</td>
<td>462</td>
</tr>
</tbody>
</table>

### Table 2  The most frequent HIO concepts identified in the Informatics literature

<table>
<thead>
<tr>
<th>HIO Concepts</th>
<th># of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>10950</td>
</tr>
<tr>
<td>Evaluate IS/IT</td>
<td>10303</td>
</tr>
<tr>
<td>Networking</td>
<td>8782</td>
</tr>
<tr>
<td>Evaluation</td>
<td>5981</td>
</tr>
<tr>
<td>Simulation</td>
<td>5286</td>
</tr>
<tr>
<td>Guidelines</td>
<td>2986</td>
</tr>
<tr>
<td>Decision support</td>
<td>1977</td>
</tr>
<tr>
<td>Neural networks</td>
<td>1768</td>
</tr>
<tr>
<td>Multivariate</td>
<td>1592</td>
</tr>
<tr>
<td>Primary care</td>
<td>1431</td>
</tr>
<tr>
<td>Study design</td>
<td>1412</td>
</tr>
<tr>
<td>Security</td>
<td>1379</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>1060</td>
</tr>
<tr>
<td>User interface</td>
<td>1009</td>
</tr>
</tbody>
</table>

### Table 3  SNOMED CT body sites covered in the Biomedical Informatics literature listed by the number of concepts identified.

<table>
<thead>
<tr>
<th>Disorder by body site</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>362965005 Disorder of body system</td>
<td>7185</td>
</tr>
<tr>
<td>399902003 Disorder of body cavity</td>
<td>4225</td>
</tr>
<tr>
<td>128121009 Disorder of trunk</td>
<td>3881</td>
</tr>
<tr>
<td>49601007 Disorder of cardiovascular system</td>
<td>2402</td>
</tr>
<tr>
<td>127331007 Neoplasm by body site</td>
<td>1693</td>
</tr>
<tr>
<td>49483002 Disorder of mediastinum</td>
<td>1643</td>
</tr>
<tr>
<td>118934005 Disorder of head</td>
<td>1524</td>
</tr>
<tr>
<td>118940003 Disorder of nervous system</td>
<td>1387</td>
</tr>
<tr>
<td>53619000 Disorder of digestive system</td>
<td>1158</td>
</tr>
<tr>
<td>928000 Disorder of musculoskeletal system</td>
<td>979</td>
</tr>
<tr>
<td>128598002 Disorder of integument</td>
<td>864</td>
</tr>
<tr>
<td>123397009 Injury of anatomical site</td>
<td>823</td>
</tr>
<tr>
<td>363170005 Inflammation of specific body structures or tissue</td>
<td>776</td>
</tr>
<tr>
<td>363171009 Inflammation of specific body systems</td>
<td>762</td>
</tr>
<tr>
<td>42030000 Disorder of the genitourinary system</td>
<td>732</td>
</tr>
<tr>
<td>50043002 Disorder of respiratory system</td>
<td>722</td>
</tr>
<tr>
<td>301810000 Infection by site</td>
<td>672</td>
</tr>
<tr>
<td>128605003 Disorder of extremity</td>
<td>630</td>
</tr>
<tr>
<td>19660004 Disorder of soft tissue</td>
<td>597</td>
</tr>
<tr>
<td>362968007 Disorder of reproductive system</td>
<td>539</td>
</tr>
<tr>
<td>362969004 Disorder of endocrine system</td>
<td>530</td>
</tr>
<tr>
<td>129565002 Disorder of skeletal AND/OR smooth muscle</td>
<td>508</td>
</tr>
<tr>
<td>105969002 Disorder of connective tissue</td>
<td>444</td>
</tr>
<tr>
<td>128127008 Visual system disorder</td>
<td>431</td>
</tr>
<tr>
<td>79604008 Disorder of breast</td>
<td>317</td>
</tr>
<tr>
<td>232208008 Ear, nose and throat disorder</td>
<td>269</td>
</tr>
<tr>
<td>362966006 Disorder of auditory system</td>
<td>209</td>
</tr>
<tr>
<td>128983005 Non-human disorder by body site</td>
<td>183</td>
</tr>
<tr>
<td>33308003 Disorder of back</td>
<td>168</td>
</tr>
<tr>
<td>414030009 Disorder of immune structure</td>
<td>163</td>
</tr>
</tbody>
</table>
is the content which defines the articles that were not indexed by our merged IMIA-AMIA Ontology, which limits the immediate applicability of our work.

The HIO requires more work to create formal definitions for the concepts that AMIA and IMIA determined to define our specialty and we need to be able to connect this ontology to an upper level ontology such as the Basic Formal Ontology [13].

Future research should include a failure analysis. One method would be to identify the MeSH terms in the articles that were not identified and see where commonalities are found across articles. Then hold a series of expert reviews to determine relevance.

The authors suggest that further research is needed to develop validated automated methods for populating the Health

Table 3  Continued

<table>
<thead>
<tr>
<th>SNOMED CT body sites and systems not identified in the 27,000 titles and abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>123646008 Disorder by body site</td>
</tr>
<tr>
<td>302918009 Disorder of stoma</td>
</tr>
<tr>
<td>302918009 Disorder of hematopoietic structure</td>
</tr>
<tr>
<td>302918009 Disorder of face and neck congenital anomalies</td>
</tr>
<tr>
<td>302918009 Hereditary disorder by system</td>
</tr>
<tr>
<td>302918009 Disorder of hemostatic system</td>
</tr>
<tr>
<td>302918009 Congenital abnormality of lower limb and pelvic girdle</td>
</tr>
<tr>
<td>302918009 Disorder of lymphatic system</td>
</tr>
<tr>
<td>302918009 Fetal disorder</td>
</tr>
<tr>
<td>302918009 Chromosomal disorder</td>
</tr>
<tr>
<td>302918009 Disorder of body wall</td>
</tr>
<tr>
<td>302918009 Disorder of product of conception</td>
</tr>
<tr>
<td>302918009 Allergic disorder by body site affected</td>
</tr>
<tr>
<td>302918009 Complication of transplanted organ</td>
</tr>
</tbody>
</table>

Table 4  SNOMED CT body sites and systems not identified in the 27,000 titles and abstracts

<table>
<thead>
<tr>
<th>SNOMED CT body sites and systems not identified in the 27,000 titles and abstracts</th>
</tr>
</thead>
<tbody>
<tr>
<td>123946008 Disorder by body site</td>
</tr>
<tr>
<td>302918009 Disorder of stoma</td>
</tr>
<tr>
<td>302918009 Disorder of hematopoietic structure</td>
</tr>
<tr>
<td>302918009 Disorder of face and neck congenital anomalies</td>
</tr>
<tr>
<td>302918009 Hereditary disorder by system</td>
</tr>
<tr>
<td>302918009 Disorder of hemostatic system</td>
</tr>
<tr>
<td>302918009 Congenital abnormality of lower limb and pelvic girdle</td>
</tr>
<tr>
<td>302918009 Disorder of lymphatic system</td>
</tr>
<tr>
<td>302918009 Fetal disorder</td>
</tr>
<tr>
<td>302918009 Chromosomal disorder</td>
</tr>
<tr>
<td>302918009 Disorder of body wall</td>
</tr>
<tr>
<td>302918009 Disorder of product of conception</td>
</tr>
<tr>
<td>302918009 Allergic disorder by body site affected</td>
</tr>
<tr>
<td>302918009 Complication of transplanted organ</td>
</tr>
</tbody>
</table>
Informatics Ontology using concepts identified in the Biomedical Informatics literature. This could be facilitated by using the context created by the semantics already contained in the HI Ontology. Once expanded, formal definitions should be added and the classification then tested both for its ability to represent new literature and also it should be subjected to a consensus review by the members of the specialty of Biomedical Informatics.

We also see this as a catalyst to continue the debate regarding defining the depth and breadth of the field of Biomedical Informatics including how we should go about expanding the knowledge base of our discipline. We relish this debate and look forward to an ongoing dialog toward synergistically improving our discipline’s self-understanding and self-awareness.

6. Conclusions
It is important for individuals in the field to understand the depth and the breadth of the specialty of Biomedical Informatics. These results show that in addition to a consensus driven top-down method of defining our specialty, there is value in looking at the Biomedical Informatics literature to provide broader definition of the concepts needed to define our specialty.

References


