MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-728

What Do They Mean by "Health Informatics"? Health Informations Posts Compared to Program Standards

Josette F Jones, RN, PhD[a], Enming Zhang, MS[a], Anand Kulanthaivel, MIS[a] , Shilpa Katta[a]

[a] *Department of BioHealth Informatics, Indiana University, Indianapolis, Indiana, USA*

Abstract

There is a lack of alignment between and within the competencies and skills required by health informatics (HI)-related jobs and those present in academic curriculum frameworks. This study uses computational topic modeling for gap analysis of career needs vs. curriculum objectives. The seven AMIA-CAHIIM-accepted core knowledge domains were used to categorize a corpus of HI-related job postings (N=475) from a major United States-based job posting website. Computational modeling-generated topics were created and then compared and matched to the seven core knowledge domains. The HI-defining core domain, representing the intersection of health, technology and social/behavioral sciences matched only 45.9% of job posting content. Therefore, the authors suggest that bidirectional communication between academia and industry is needed in order to better align educational objectives to the demands of the job market.

Keywords:

Employment (D004651), Curriculum (D003479), Medical Informatics (D008490)

Introduction

The field of health informatics (HI) is a relatively new area of study and practice closely related to, and sometimes used interchangeable with *clinical informatics*, *medical informatics*, and *biomedical informatics*. Currently only the American Medical Informatics Association (AMIA) and the Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) provide a curriculum competency framework [1] describing the minimum knowledge and skills graduates of a HI program must attain. No governmental entity provides an official labor classification for professional in health informatics nor is there an agreement on the job titles.

The question arises if the competencies set forth by the curriculum framework are in harmony with the knowledge and skills required by the job market; Kulikowski et al [2] brought in one of the broadest definitions of health informatics then coined as *biomedical informatics (BMI)*. In this 2012 white paper, the authors proposed that core competencies in BMI would include basic biomedical science, information technology, computing, professionalism, and knowledge of the social science of information usage. More recently, Fridsma [3] published a viewpoint in the *Journal of the American Medical Informatics Association (JAMIA)*, using the concept of HI and no longer BMI to refer to the domain. HI is described as "intersections creating a continuum" ([3], p. 855) between various knowledge domains of application including healthcare, biosciences, computing, and social sciences and ranging from a disease management to public health applications and research.

The broadly stated competencies described in the 2012 AMIA white paper [2] were not yet in a form usable for formal accreditation processes. Since 2016, the AMIA Accreditation Committee (AAC) has been reframing and redefining through an iterative process the curriculum requirements as graduate outcomes. The resulting revisions [1] set forth 10 foundational domains, each with accompanying knowledge, skills, and attitudes necessary to succeed as health informatics professionals in an ever-changing job market. The three base domains are: Health ("F1"; biomedical and health sciences), Information Science & Technology ("F2"; methods and technologies for storage and exchange of information), and Social & Behavioral Science ("F3"; psychology, sociology, organizational behavior).

Three second-level intersected domains are formed by the joining of any two of the above base domains: F1 and F2 combine to form Health Information Science & Technology (F4), F2 and F3 combine to form Human Factors & Sociotechnical Systems (F5), and F1 and F3 combine to form the Social & Behavioral Aspects of Health (F6).

The unique core of health informatics is at the intersection of F1, F2, and F3; the intersection is known as "Social, Behavioral, and Information Science and Technology Applied to Health" (F7), and its specific knowledge, skills and attitudes are considered the defining and differentiating foundation HI program. [1]

The AAC added three domains: Professionalism, leadership, and interprofessional collaborative practice. The relationships and intersections among the original seven domains are visualized in Figure 1. The AAC's three added domains, although important, are not considered in this analysis as these domains are considered intrinsically required by most jobs, regardless of field. The foundational domains of HI education are depicted in figure 1.

Yet still the question remains: do the knowledge, skills and attitudes set forth in the most recent recommendations [1] cover the demands of the job market?

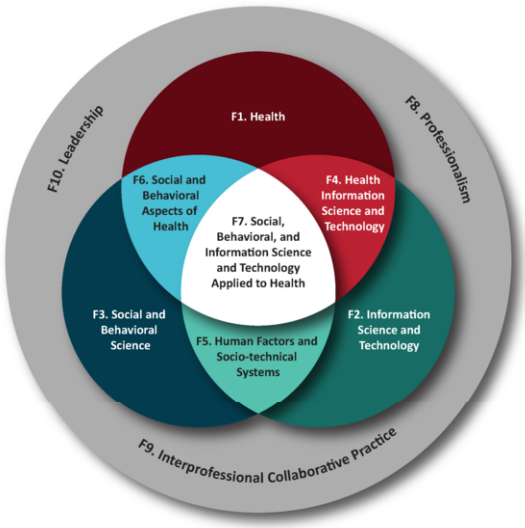

Figure 1. Venn Diagram of the Competency Domains Proposed by the AAC, 2016 [1].

Objective

Due to the inconsistencies in job descriptions and the recent revisions in HI program outcomes and related competencies, the need arises to perform a gap analysis between the knowledge and skill domains identified for the HI curriculum and actual posted job requirements.

Methods

NLP: Computational Topic Modeling

A convenience sample of US-based job postings made as of mid-November 2016 from the website Indeed.com [4] was searched and mined using the Publisher Toolkit's application programming interface [5] connected to a proprietary Python-based script. To extract the HI-related postings, the terms "*health informatics", "clinical informatics",* and "*medical informatics"* were used to filter the resulting corpus. The resulting filtered data were then organizing into individual documents, each document containing one job posting.

MALLET (Machine Language Learning Toolkit) [6] is an open-source topic modeling software tool that utilizes an algorithm that computes baskets-of-words (named topics) that frequently occur across a set of documents. MALLET assigns individual weights to each basket-of-words within each document. For this study, MALLET output was exhausted at 30 topics co-occurring across these 475 job-posting documents.

Qualitative Second-Level Matching

Following MALLET categorization, *second level matching* was performed: Two HI graduate students independently assigned one of the seven core AMIA/CAHIIM knowledge domains [1, 2] to each MALLET-generated keyword basket. The inter-rater agreement was calculated at this stage (23/30 topics; 76.7%). Disagreement was resolved by mutual adjudication. A similar process was used for labeling spurious topics such as those referring to marketing publicity.

The topic weights were then summed to determine the weight of each individual domain from F1 through F7, including the spurious topics, within each job posting. For example, if a job matched to the three MALLET generated topics for foundational domain F2 at strengths 0.0528 (software development), 0.000140 (data science), and 0.00044 (software application) respectively, the overall weight for domain F2 would be 0.0534.

Binary Model Generation

The matrix was simplified into a binary matrix – potential combinations of the foundational domains excluding the nucleus of HI - by only including relative job-domain strengths that exceeded 0.242 (equivalent to one standard deviation above the median). These strengths were coded as "1", while lower associations were coded as "0".

Reclaiming Missed Postings

In order to verify that all potential HI job postings were extracted from the original list, we analyzed those jobs that did not match F7 in the binary model but matched entirely one of the combinations below:

- F1 + F2+ F3
- F3 + F4
- F1 + F5
- F2 + F6
- F5 + F6
- F4 + F5
- F6 + F7

Results

NLP: Computational Topic Modeling

The retrieval procedure (downloading from Indeed.com) yielded 475 job postings that matched the search query.
MALLET determined 30 topics. In order to determine the relative weight of each domain across the corpus, the document-topic strength file generated by the software was analyzed.

An example of the data job-topic data generated by MALLET is seen in the following table:

Topic	T00	T01	T02
Job 0	0.0001	0.0624	0.0528
Job 1	0.0001	0.1594	0.0002
Job 2	0.0000	0.0004	0.0001

Table 1. Document (Job posting)-to-Topic Weights.

Qualitative Second-Level Matching

Each topic was qualitatively assigned a domain by the panel of two graduate assistants. A list of topics, along with the qualitatively assigned knowledge domains, is available in Appendix A.

During the rating of the MALLET-generated topics, the graduate assistants also excluded by consensus any topics they determined to be unrelated to actual core knowledge domains. Specifically, 8 out of 30 topics generated were deemed by consensus to be spurious as shown in the table below.

Reason deemed spurious	# of Topics
Exclusively advertised the employer	5
Exclusively described employee physical characteristics (physical labor capabilities) and not employee competencies	2
Consisted entirely of mandatory legal language (equal opportunity statement, etc.)	1

Table 2. Spurious Topics and Frequencies.

The topics were then collated into cumulative strengths that represented the strength of each *domain* across the postings. The median of all job-domain strengths was found to be 0.024; the standard deviation of job-domain strengths was 0.218. An example of cumulative strengths by domain for job postings 0, 1, 2, and 3 across F4, F5, and F6 is below.

Job	F4	F5	F6
0	0.000	0.199	0.594
1	0.000	0.027	0.227
2	0.000	0.001	0.997
3	0.001	0.002	0.989

Table 3. Job Posting Cumulative Strength by Domain, Examples.

Binary Model Generated

The full binary job-domain relationship table was generated; a a truncated example of the results (derived from Table 3's domain strengths) is in the below table.

Job	F4	F5	F6
0	0	0	1
1	0	0	0
2	0	0	1
3	0	0	1

Table 4. Binary Representation of Job Posting Cumulative Strengths by Domain.

The binary model also revealed the following match rates for each domain:

	Postings Matched	
Domain	**N**	**%**
F1	39	8.2%
F2	44	9.3%
F3	12	2.5%
F4	29	6.1%
F5	43	9.1%
F6	376	79.2%
F7	40	8.4%
(Spurious Topics)	173	36.4%

Table 5. Binary Results for Matching of Job Postings to Topics Related to Each Competency Domain.

The domain (F7) considered the unique core of HI matched 8.4% postings at an overall strength of 0.242 or higher. 36.4% of postings matched spurious topics, although only 43 (8.8% of corpus total) job postings matched *only* the spurious topic set. The most job descriptions (79.2%) match the competencies expressed in foundational domain F6: social and behavioral aspects of health systems. One job posting (0.2%) did not match any of the topics at a binary threshold of 0.242, meaning that the total adjusted match rate for *any* non-spurious topic was 90.9% (N=433).

Reclaiming Missed Postings

The lack of postings (8.4%) that initially matched the F7 domain is of concern. This deficit, however, was amended by *reclaiming* postings that missed classification as F7. In order to increase the match rate for job postings that potentially related to the F7 domain, combinations of co-matches between topics that together would make up F7 were sought, increasing the F7 match rate from 8.4% (N=40) to 24.2% (N=115).

Domain Combination	N	% of corpus
F1 + F2+ F3	0	0.0%
F3 + F4	0	0.0%
F1 + F5	1	0.2 %
F2 + F6	33	7.0%
F5 + F6	23	4.8%
F4 + F5	0	0.0%
F6 + F7	18	3.8%
Recovered Total	75	15.8%
(F7 initial)	40	8.4%
Sum: Total possible matches	**115**	**24.2%**

Table 6. Job Postings Recovered by Combinations of Domains that Could Represent F7.

Discussion

The analysis revealed certain matches between the curriculum competencies and the skills and knowledge desired in 475 real life job postings on *health informatics* and/or related terms.

The least frequently matched domain was Social & Behavioral Sciences (F3), matching 2.5% of job postings. This finding implies that employers looking for HI-related professionals may seek fewer individuals who solely have social-behavioral science competency. The most popular competency domain correlated to these job postings, on the other hand, was observed to be Social & Behavioral Aspects of Health (F6), which matched 79.2% of all job postings in the binary model, indicating that combined competencies bridging Social & Behavioral Sciences with Health Sciences are in relatively high demand.

Match Rate for F7 – Social, Behavioral and Information Science and Technology Applied to Health

The reclaimed postings, while they only partially match to the core of HI [1,2], will still be considered related to Health Informatics.

The F7 match rate (maximum of 24.2% even after reclamation) remains lower than the F6 match rate (79.2%) suggesting that there exist many *health informatics*-related jobs that *employers describe as* requiriung knowledge of social sciences and health while requiring little to no technology competency.

Use of Qualitative (Second Level) Matching

It also follows that qualitative *matching* analysis to interpret the *keyword baskets* generated by NLP software such as MALLET may have its pitfalls due to potential bias on behalf of the raters. Nonetheless, qualitative secondary matching has been shown of use in health related applications including coding medical concepts from clinical free text entries [7] and classification-annotation of mentions of pharmaceutical treatments [8] the authors of both of these studies implore that qualitative analysis is required following computational NLP.

Limitations: MALLET Analysis

It is also noted that there is required a future qualitative analysis in terms of quality assurance of the MALLET-generated results. In the future, other analyses should be undertaken in order to study this and similar corpora: Further study may include synonym-based (synset) literal term searching, as well as qualitative analysis of a reduced corpus in order to perform an expert categorization of job postings.

Limitations: Spurious Topics & Posting Bias

The presence of what the authors term *spurious topics* (those that exclusively advertised the employer, location, or legal requirements) must also be addressed, as these spurious topics matched to 36.4% of all job postings in the binary model. In fact, 8.8% of all job postings matched *only* the spurious topics.

The corpus that was searched and processed may therefore represent a biased sample. Specifically, it likely represents the wording chosen by the writers of the job postings. While the postings did contain information about the core knowledge domains required to satisfy job demands, it is also observed via the strength and content of spurious topics that significant amounts of advertising and legal information was present in the job postings. Furthermore, if there is any inaccuracy or inexperience on behalf of those writing the job postings, such error cannot be taken into account by the study design at hand.

Limitation: Incomplete Coverage of Domains

The authors also note that a limitation exists in the coverage of all competency domains and their expressions of knowledge, skils and attitudes.. While the CAHIIM-AMIA competency classification includes 10 domains, only the 7 domains from the overlapping Venn diagram were covered in this proof of concept study. Future studies will need to cover all 10 competency domains in order to better match job requirements to curriculum competencies.

Conclusions

The study reported was able to create a connection between a major upcoming curriculum competency domains and real life job postings for health informatics and related areas. Computational topic modeling/NLP followed by qualitative (*second level*) consensus analysis has here shown the ability to process a large corpora (N = 475) of job postings and assign academically-meaningful topics to 433 (90.9%) of them.

Therefore, the authors recommend that NLP via computational topic modeling, followed by qualitative second-level analysis, be used to analyze further corpora of job postings in order to match them to curriculum frameworks. Such matching analysis is likely to reveal the similarities and differences between curricular needs and the job market; this information can be used to tailor curriculum frameworks as well as employer job postings.

Acknowledgements

The authors would like to acknowledge the AMIA Accreditation Committee for their revisions of the foundational domains of Health Informatics.

References

[1] Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM) & American Medical Informatics Association (AMIA). AMIA Accreditation Committee : Draft For Public Comment, 2016.

[2] Kulikowski C, Shortliffe E, Currie L, Ekin P, et al. AMIA board white paper : Definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *J Am Med Inform Assoc* (2012), 931-938.

[3] Fridsma DB. The scope of health informatics and the Advanced Health Informatics Certification, *J Am Med Inform Assoc* (2016), 855-856.

[4] Indeed.com, Inc. Job Search. (Website), n.d.-2016. Retrieved from http://www.indeed.com/

[5] Indeed.com, Inc. Publisher API. (Online Software Tool), 2016. Retrieved from http://www.indeed.com/publisher

[6] McCallum AK et al. MALLET : A Machine Learning for Language Toolkit. (Software Tool). Amherst, MA : University of Massachusetts, 2002. Retrieved from http://mallet.cs.umass.edu/

[7] Lin CH, Wu NY, Lai WS, Liou DM. Comparison of a semi-automatic annotation tool and a natural language processing application for the generation of clinical statement entries. *J Am Med Inform Assoc* **22** (2015), 132-142.

[8] Merabti T, Abdoune H, Letord C, Sakji S, Joubert M, Darmoni S. Mapping the ATC classification to UMLS metathesaurus: Some pragmatic applications, *Stud Health Technol Inform* (2011), 206-213.

Address for correspondence

Corresponding Author: Dr. Josette Jones.
Email: jofjones@iupui.edu.
Mailing Address: 719 Indiana Avenue, Walker Plaza #319, Indianapolis, Indiana USA 46202.

Appendix A: Word Baskets & Topics with qualitative interpretation

Topic No.	Topic/Word Basket	Interpretation
T00	nursing colorado library school pharmacy city campus kansas hays researchers board state anschutz denver regional schools resources hit access umkc	Spurious
T01	health research informatics public sciences programs school position care medicine policy services education center program demonstrated population social staff medical	F6
T02	experience **development software** web team strong years programming java informatics technologies boston company equivalent developer applications gns cloud performance science	F2
T03	skills work experience ability knowledge information health management communication working team position informatics related required degree project provide projects including	F6
T04	business sales customer account market marketing develop product opportunities products strategies healthcare create drive strategy experience industry accounts key customers	F6
T05	research center data science bioinformatics computing computational biomedical scientific medicine school cancer biology texas collaborative expected genomics candidate postdoctoral university	F4
T06	management health information development technology support planning leadership administration including policies business operations experience initiatives program strategic objectives plans programs	F5
T07	infrastructure including pharmacy alliance tobacco application architecture results benefits develop network medicine university compute cancer monitor pcori build technologies control	F5
T08	usf research position clinical florida universities tampa public opportunity working cover equal employees institute south system apply resume click top	Spurious
T09	develops maintains related management performs quality requirements ensures program works participates functions utilization serves reports procedures leads assists demonstrates services	F3
T10	status employment opportunity equal gender national disability protected sexual race color religion age origin veteran orientation identity applicants employer law	Spurious
T11	college georgia education students student community campus faculty school degrees university academic technology duties nursing coastal advising studies experience ccga	Spurious
T12	health information clinical informatics care data technology nursing science practice systems develop design applications administration management implement assist staff computer	F5
T13	data public health disease healthcare american surveillance heart association epidemiology prevention opportunity resume division fellowship duties city stroke impact control	F1
T14	philips health job sales company act care opportunity solutions home contact technology consumer application process applicants title employer clinical protected	Spurious
T15	university faculty teaching department information position program students graduate candidates science research qualifications applications professor informatics application professional college include	Spurious
T16	project manager united healthcare management experience specific ability electronic account skills implementation members role primary patients job diagnosis knowledge team	F6
T17	data analysis health experience business statistical reporting analytic analyst tools complex required research reports preferred years analytics requirements including database	F7
T18	healthcare business team solutions data clients work services teams project product organization client stakeholders lead analytics care ensure processes understanding	F6
T19	quality clinical health measures experience federal measure measurement programs review working project care research nqf cms state performance public agencies	F6
T20	healthcare sales digital quality customer clinical solutions business product leadership including customers strong products global company imaging willingness ambulatory technical	F6
T21	quality care health data improvement reporting healthcare performance activities provider improve meaningful providers reports medicare patient measures analysis national analytics	F6
T22	coding health information documentation icd required management knowledge registered program rhia ccs certification rhit education accurate medical director reimbursement codes	F6
T23	medical group payer building rules youâ€™ll ideas improve finance growing performance make review join relationship industry unitedhealth payers culture insurance	Spurious
T24	ability essential physical job perform employee functions demands pharmacy time required vision occasionally individuals including reasonable disabilities pounds telephone made	Spurious
T25	patient service patients department ensure field position safety policies medical reports staff daily procedures equivalent mount supervisor competencies customer current	F6
T26	**data** learning required machine analytics knowledge solutions mining predictive modeling language group decision iet models janssen processing statistics **sciences** scientist	F2
T27	clinical informatics system information systems workflow ehr user technology workflows michigan training end physician users specialist leadership patient data setting	F4
T28	**systems** software system technical application computer requirements support design development epic **application**s testing technology documentation solutions users services maintain user	F2
T29	clinical medical experience care preferred staff education quality patient practice required years ensure healthcare improvement duties physicians assigned minimum training	F1