A Multi-Stage Curriculum Development Model to Address Knowledge Gaps Between Academia and Industry

Heidi Douglass
Department of BioHealth Informatics
Indiana University
Indianapolis, United States
hgdougla@iu.edu

Gary Schwebach, J.D., D.B.A
Department of BioHealth Informatics
Indiana University
Indianapolis, United States
gschweba@iu.edu

***Abstract*— This innovate practice research is a work in Progress. Curriculum for technology programs must remain current with the recent developments and present needs of industry to ensure their continued success. Identifying and incorporating the newer developments requires a knowledge of those things that are important to the commercial users of technology. We have proposed and are testing a methodology that can easily allow educational programs to remain current using natural language processing followed up with confirmatory surveys. Our research used employment postings and website descriptions of ongoing research/ product development from technology companies to identify current interests. This data provided a window into where companies have invested their resources for future growth, which we used to understand where their technology needs were. The findings will then be used to determine what changes a program should make in their curriculum to address these needs.**

The information from these different sources was collected from the websites of selected technology companies as well as from technology related job listings. We then programmatically determined the ten most frequent words found in each subsection of the job listings search parameters. A correlation analysis for each word was obtained to give a context to the words in the resulting corpora. Our next step in developing this process was to extract more focused topics in our data from the correlations using focus topics extracted from the information available in the technology companies' sites.

The process used to focus the topics in our data was started by capturing research and services related to health informatics and bioscience informatics. These research and services were obtained from a sample of 34 local and 39 national companies. The University of Massachusetts' (Amherst) free and open source machine learning toolkit, MALLET, was used to discover the topics present in the datasets via Latent Dirichlet Allocation (LDA) and then compare and interpret each topic present in the job posting frequent word correlations. This allowed us to determine more specific informatic subjects of interest along with the importance of these subjects to industry.

Immediately following the analysis, we will incorporate our results into a survey, which will be sent to the companies that we identified in our data search. This will be used to validate our findings as well as obtain more specific information on their needs related to the curriculum. This second component will allow us to obtain more actionable information for making and incorporating new information and materials into the curriculum.

Keywords*— *Natural Language Processing, Curriculum Development

I. Introduction

Informatics and technical innovation have rapidly developed in biotechnology and healthcare technology industries in the United States over the last two decades [1, 2, 3]. A problematic gap in the ability to use these innovations that originated in academic and industry settings, including smaller startup biotechnology companies and large pharmaceutical companies, has led to a need for increased collaboration. [4 - 9]. Globally, biotech and pharmaceutical industries are seeking to collaborate with academic researchers to acquire the necessary information and expertise to fully utilize these innovations using curriculum and collaborations that address these needs. [5, 6, 9, 10]. However, published research pertaining to the transfer of innovation from academia to industry knowledge over the last five to ten years specific to the United States was not found. We put forth that bridging this gap can be partly addressed by a better understanding the needs of industry when designing curriculum in computer science, informatics and engineering courses.

There is an obvious concern that academic institutions should not become "trade schools" for industry and thus lose their mission of exploring and expanding the body of knowledge in each respective area. This research is premised on the belief that situating academic learning and research in actual problems, especially in the STEM disciplines, provides a good learning environment while simultaneously allowing for theoretical development of new knowledge. This will motivate students to perform at higher levels due to the future possibilities after graduation while also increasing the opportunities for translational research based on theoretical development.

A. Background

Our research originated when the Department of Bio-Health Informatics (DBHI) at the Indiana University School of Informatics and Computing, Indianapolis was approached by local biotech companies to develop course offerings specifically to address this lack of knowledge. However, because of the financial and time constraints of enrollment for advanced degrees, an appropriate short-term solution to bridge this knowledge gap was to create course offerings that provide specialized training for employees. The subject matter curriculum for these course offerings needed to be pertinent to the local biotechnology and healthcare technology industry. However, as we began this research, we determined that the model we were developing would have application to all technology education. Therefore, we are presenting our model to help other institutions develop curriculum that is relevant and meaningful to the industries that will seek their graduates.

Our working model to develop the curriculum for these course offerings involved multiple stages. First, a panel of subject matter experts were assembled to understand the broader needs of the bio health technology industry. The next stage in the process provided a way to assess the presence of the reported biotechnological demands using Natural Language Processing (NLP) on state and national job listings and biopharmaceutical as wells as healthcare technology research and services. It is those results that are presented in this article. Our final stage in determining the curriculum offerings will use a questionnaire that will be administered to local biotechnology and healthcare technology businesses to further refine and validate our findings and conclusions from the NLP analysis for actual curriculum development.

B. Implementation of Natural Language Processing

Natural Language Processing has been used for multiple applications in the research setting that can translate to curriculum development. Words and phrases are identified in large bodies of literature to reveal trends or changes in content and even concepts [13-16]. These techniques have previously been employed in the development and evaluation of medical education curriculum [13]. Using these capabilities, the informatics and technology training needs of local industry by observing what they were actively seeking for their commercial success. Initially, we utilized the information extraction capability of NLP to obtain the most frequent words found in job listings that pertained to informatics-based skills communicated by healthcare analytic and biopharmaceutical /biotechnology businesses. We then examined these words using correlational analysis to understand the context of the technological demands of industry. Because this context was broad, more focused topics in the Bio-Health technology sectors' research and services were obtained to interpret and focus the demands. These findings will then be confirmed by asking the local business leaders how well our model has captured their needs by using a survey methodology.

II. Methods

The process in understanding the healthcare technology and biotechnology demands for targeted curriculum development involved a multistep process An advisory board of subject matter experts (SMEs) from the biotechnology and healthcare technology industry provided us with a direction and a description of potential needs that these business sectors had identified.

Using these findings, we developed our methodology then applied it to the state level for further refinement and validation. Our location in Indiana guided this part of our development cycle. We then applied it to the national level in the United States. The resulting model was based in the needs of biotechnology and healthcare technology at both the state and national levels (Figure 1).

To investigate the informatics demands within the state as well as national healthcare technology and biopharmaceutical companies, data describing the needs of these companies were obtained via job postings. These job postings were scraped from Indeed.com, a national job posting website that provides listings with a high specificity to these companies. Search parameters that returned job postings requesting talent to have or perform informatics-based tasks were also optimized. The resulting search parameters used on Indeed.com included full, part-time, or internship jobs in the state of Indiana and the United States pertaining to biotechnology, research scientist, scientist, healthcare analytics, and healthcare information technology. The returned jobs were scraped, downloaded, and analyzed. The Google Chrome extension Web Scraper API extracted the job data from each search and exported these as a CSV file. Subsequently, the downloaded data were divided into two categories using the original search descriptors. Biotechnology, research scientist, and scientist were grouped into biopharmaceutical jobs, and healthcare analytics and healthcare information technology were grouped into healthcare technology jobs. The data collection was conducted from November 2019 through February 2020. The biopharmaceutical jobs totaled 348 at the state level and 452 at the national level. The national healthcare technology jobs obtained totaled 246 and the state jobs totaled 201. These data were analyzed and interpreted according to their respective regional categories.

A. Data Processing

Each category of data was formed into a corpus with Ingo Feinerer's "tm" package for R's "Corpus" function. The corpora for the job listings were transformed in six steps. Whitespace was stripped, letters were converted to lowercase, and all numbers and punctuation as well as common words in the English language were removed [17]. Subsequently, words that were not pertinent to the companies' informatics needs such as retirement benefits, were also removed. The processed corpus was transformed into a term document matrix with the "tm" package previously referenced. These data were sorted by word frequency and the transformed corpus visualized with a word cloud.

After obtaining the ten most frequent words for each listed category, words correlated with these words in the data were obtained to determine the relevant contexts from which the words were used in the job postings. The word correlations with

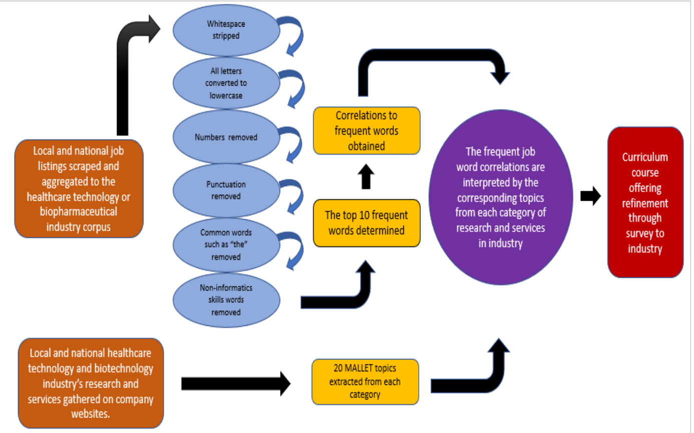

Local and national job listings scraped and aggregated to the healthcare technology or biopharmaceutical industry corpus
Whitespace stripped
All letters converted to lowercase
Numbers removed
Punctuation removed
Common words such as "the" removed
Non-informatics skills words removed
Correlations to frequent words obtained
The top 10 frequent words determined
The frequent job word correlations are interpreted by the corresponding topics from each category of research and services in industry
Curriculum course offering refinement through survey to industry
Local and national healthcare technology and biotechnology industry's research and services gathered on company websites.
20 MALLET topics extracted from each category

Fig. 1 *The workflow for the development of curriculum based in industry's needs*

a Pearson correlation value of 0.4 or greater as determined by "findAssoc" method in the previously referenced "tm" package for R were examined for relevance to the informatic or analytical aspect of a business's needs through visual inspection. Correlations that described a company's adherence to state and federal laws, employee benefits, or contractual agreements required for the hiring process were disregarded.

To further understand the applicability of the frequent words and their correlations, a comparison was made to topics extracted from biopharmaceutical and healthcare technology companies' research and services descriptions. 34 companies' descriptions were obtained from the top 100 life science companies (by employment) in the state of Indiana as reported by BioCrossroads® [12] that offered a description of their current research or services for agricultural development, drugs and pharmaceuticals, health information technology, or research, testing, and medical laboratories available on the company websites. Implementing the same process at the national level, 39 high-grossing biopharmaceutical and healthcare research/technology companies' available research and services were obtained. The health information technology companies were analyzed separately from the biopharmaceutical companies to maintain an understanding of the research and services that are specific to each field.

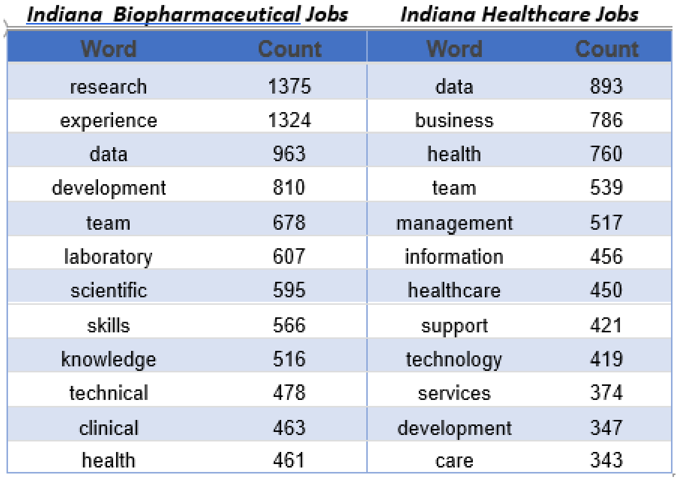
A Latent Dirichlet Allocation (LDA) method utilized by the tool MALLET extracted 20 topics per state and national company research category. These topics were compared to the corresponding category of national and state job listing word correlations to refine an understanding of the informatics needs present in the listings. After the topics were refined with the domains found at the local and national level, obtained informatics topics were used to determine the curriculum development of potential course offerings

B. *Topic Discovery*

The University of Massachusetts' (Amherst) free and open source machine learning toolkit, MALLET, was used to discover the topics present in the datasets via Latent Dirichlet Allocation (LDA). 20 topics per state and national company research category were obtained. These topics were compared to the corresponding category of word correlations to refine an understanding of the informatics needs present in the listings. After the correlations were refined with the topics found at the local and national level, focused informatics topics were used to determine potential curricula.

III. RESULTS

A. *State level biopharmaceutical and healthcare jobs analysis*



Indiana Biopharmaceutical Jobs		Indiana Healthcare Jobs	
Word	Count	Word	Count
research	1375	data	893
experience	1324	business	786
data	963	health	760
development	810	team	539
team	678	management	517
laboratory	607	information	456
scientific	595	healthcare	450
skills	566	support	421
knowledge	516	technology	419
technical	478	services	374
clinical	463	development	347
health	461	care	343

Table 1 Indiana biopharmaceutical and Healthcare job listings 12 most frequent words

The analysis of the biopharmaceutical and healthcare technology job listings for the state of Indiana showed frequent words that pertain to data management and analytics such as data, technology analytics, and technical (Table 1, Fig. 2 & 3). Next, the correlated words to the most frequent words were examined.

From this we determined the state level biopharmaceutical companies' drive for software development in data acquisition and analysis to optimize product pipelines (Fig. 2). Word correlations such as architecture, applications models, and software to the frequent word experience along with innovation to the frequent word technical revealed this demand. Additionally, correlations such as metrics and query to the

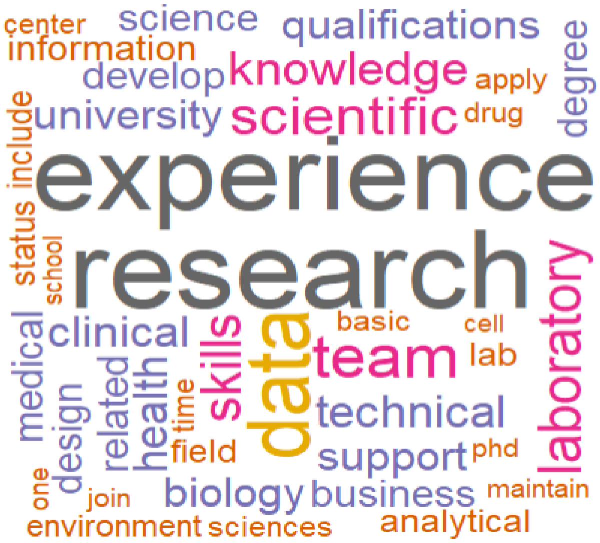


Fig. 2 Indiana biopharmaceutical job listings most frequent words

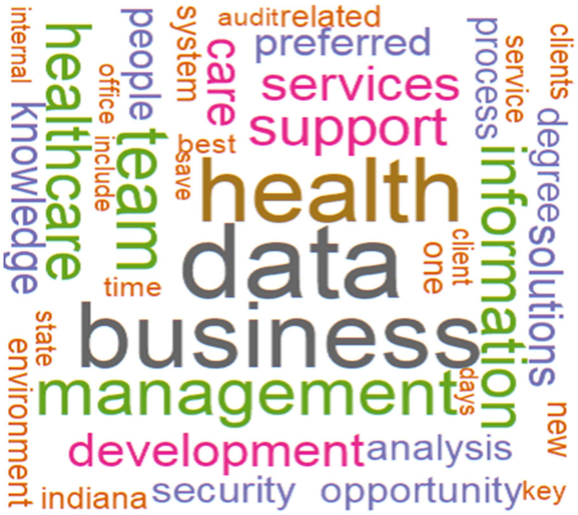


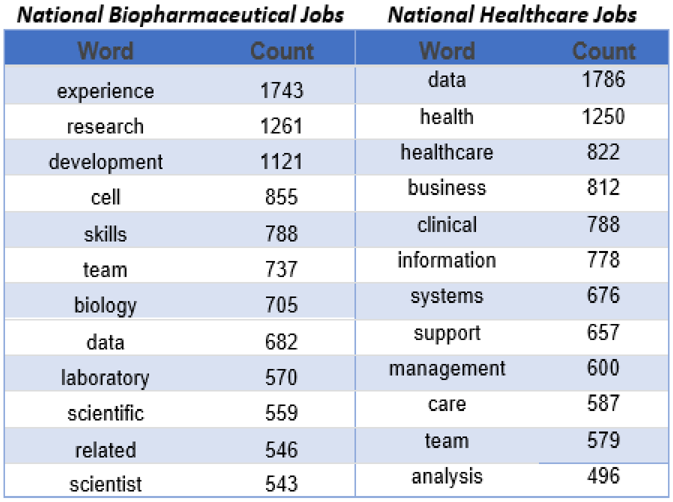
Fig. 3 Indiana healthcare technology job listings most frequent words

frequent word data and evaluation to the frequent word technical showed the industries reliance on data analytics to optimize product development.

The correlations to the healthcare technology job listings were found to focus on data security, storage, and access (Appen. B). Frequent words correlated with words in the listings such as security, cryptography, and security operations center revealed a drive for security while demand for data storage and access were understood with correlated words such as centralization, Healthcare Information Management Systems Society (HIMSS) as well as Databricks and NoSQL.

B. National biopharmaceutical and healthcare jobs analysis

A comparison of the job postings of national biopharmaceutical companies to those pertaining to national healthcare technology revealed differing informatic-based demands in these two industries. Among the top ten-word frequencies in the national biopharmaceutical companies' job postings experience, research, development, and data were the first three most frequent words (Table 2, Fig 4). From the healthcare technology category job postings, data, business, clinical, analytical, and information were among the most frequent words (Table 2, Fig 5). Examining the words correlated to these frequent words showed the biopharmaceutical postings focus on analyzing and implementing information gathered through research (Appendix C). Additionally, healthcare companies' concentration on storing data and making data driven business decisions were also found through these correlations (Appen. D).



National Biopharmaceutical Jobs		***National Healthcare Jobs***	
Word	**Count**	**Word**	**Count**
experience	1743	data	1786
research	1261	health	1250
development	1121	healthcare	822
cell	855	business	812
skills	788	clinical	788
team	737	information	778
biology	705	systems	676
data	682	support	657
laboratory	570	management	600
scientific	559	care	587
related	546	team	579
scientist	543	analysis	496

Table 2 National biopharmaceutical and healthcare technology job listings 12 most frequent words

The correlations to the job postings' frequent words for the associated biopharmaceutical companies gave insight into the data demands for analytical pipeline development (Append C). Words such as analytical and validation that are correlated to development as well as designing, engineering, develop, implement, scale, and robust correlated to the frequent word cell demonstrate aspects of this pipeline development. Other correlations to frequent words also demonstrate current research interests of the companies posting the jobs. For example, microbioreactors, transfection (CRISPR system), and clonal/ antibody assay all pertain to frequent knowledge domains in the gathered postings.

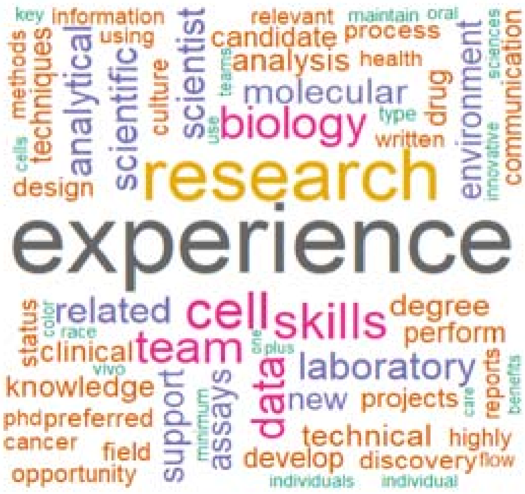


Fig. 4 National biopharmaceutical job listings most frequent words

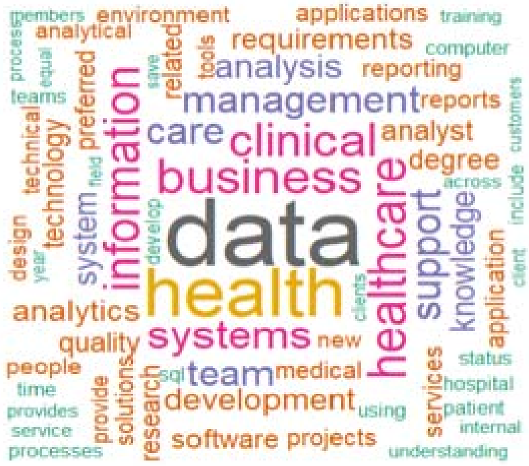


Fig. 5 National healthcare technology job listings most frequent words

Similarly, correlations in the national healthcare job postings reveal a reliance on stored and processed data to drive business decisions (Appen. D). For example, the word correlation between data to SQL, query, analysis, visualization, and tableau along with correlations between medical to database, coordinate, assessments, and screenings show the demand for data access and visualization to drive business decisions. The correlations to the word analytical for national healthcare companies add to the evidence for the demand for analytic processes with words such as AAAI (Association for the Advancement of Artificial Intelligence) and high dimensional.

C. LDA to model topics in research and services

In the next phase of our curriculum development, the general topics discovered in the frequent word correlations for state and national categories of job listings were compared to 20 topics extracted by the tool MALLET via LDA from each category of companies' research and services. These topics were used to focus the business demands described by the word correlations in order to determine possible subjects for future curricula development. A complete list of topics extracted for each category of research and services obtained is available in Appendices E – H.

At the state level, the biopharmaceutical companies most frequent topics pertained to protein functional and structural studies such as biomarker screening, monoclonal antibody studies, and cytokine release assays (Appendix E). Studies implementing genetic manipulation and observation were prevalent at the state level as well with topics covering interferon beta expression for tumor suppression, high throughput platforms, CRISPR, and gene therapy. When these topics were applied to the frequent word correlations revealing the drive for software development to acquire and analyze data along with optimizing product development, the useful topics for possible curriculum were determined to be protein identification of structure and function optimization with machine learning as well as machine learning applications to gene sequence and function identification.

When Indiana healthcare technology research and service topics were examined (Appen. F), topics specific to data access, analytics and security were identified. These topics were present in the correlations from the job listings at the state level indicating a strong demand in this sector for these topics. Specifically, topics such as electronic medical systems enrollment, optimizing healthcare testing, and analysis of stroke research, data integrity in clinical trial management, and managing patient data in analysis pipelines were drawn out from the body of research and services data. These specific topics were used to interpret the demand for data access and security found in the frequent word correlations. The topics for the development of a curriculum specific to Indiana institutions interpreted from these data were cyber security in electronic health applications and database applications with the development of a centralized platform for research analysis.

The national research and services for biopharmaceutical companies' topics returned topics that focused on therapies at the cellular level such as stem cell therapies, anticoagulant bleeding treatment studies, and immunogenicity tolerability (Appendix G). Topics particular to disease research such as multiple sclerosis, Alzheimer's, HIV, and HCV were also described by the topic extraction. Interpretation of the frequent word correlation with these specific topics yield curriculum recommendations presenting techniques in machine learning to optimize product/drug pipeline development as well as to providing robust research data analytics to those institutions who recruit students nationally.

Topic categories pertaining to disease were found in the national healthcare technology research and services data as well (Appendix H). Topics pertaining to cancer studies such as prostate cancer, adenomas, immunotherapy for tumor treatment, sequencing cancer cells, and leukemia. Diseases and pathogen topics returned were coronavirus, cardiovascular disease, and brain disorders. Additionally, optimization using technology in data collection and patient experience during treatment were returned as topics found in healthcare research. Data analytics and visualizations applied to electronic health records and research data from the previously listed diseases were the interpreted areas for curriculum relevant for academic institutions who have a primarily national recruiting goal.

a.

IV. Discussion

From our analysis, it was observed that the correlating words described a context more specific to research areas in the national subset of the job listings as compared to the state level. A possible explanation for this finding lies with the number of unique job listings' available in the national subsets. There were 104 more biopharmaceutical jobs and 45 more jobs for healthcare technology nationally compared to the state jobs that were acquired for our analysis. The smaller job availability at the state level can be addressed by scraping state jobs every 30 days until the amount of the unique national job listings are reached. The larger number of unique job listings would provide more correlations and context to the frequent words obtained for each category.

Although the Mallet topics extracted from the research and services available on companies' websites enabled a more focus understanding of the word correlations drawn from job listings, these data were drawn from only published material or links made available on the websites of the companies investigated. Because this body of data is a subset of research and services, it does not represent the complete body of research each industry was actively performing. A possible solution to the gaps in this data would be found in acquiring patent submissions or grant offerings from industry. Topics extracted from these data would then give more complete and current interpretations to the job listings' informatics demands found in the word correlations.

A. The next stage of research methodology

Our research to date provides a very good idea of the areas for curriculum development but lacks granularity in the specific course content that could be developed. The balance between providing a complete education while addressing issues relevant to industrial needs, especially in innovative areas, is imperative to achieving the needs and goals of the academic and the business communities. The initial stage of our research goes a long way to identify the balance, but curriculum development requires more.

We identified that the best way to accomplish this joint goal was to conduct further research with industry. We determined that a survey methodology would provide us with the means to validate our previous findings while further refining our information for the curriculum development.

We are in the process of developing a short survey instrument to collect additional information from the companies that we included in the initial stages. This survey instrument will be designed to collect multiple data points that are relevant to curriculum development. In particular, we will measure two elements that have a direct value in curriculum development, the need for specific topics to be included in the curriculum and the relative perceived value of each topic to the business. We will use a gap analysis methodology to determine the first and a value allocation methodology for the second.

The gap analysis model measures the difference between importance and satisfaction with the current educational opportunities in specific informatics skills. We will use the NLP results, correlation findings and the LDA results to develop a set of offerings, which we will incorporate these into our survey instrument. Example survey questions for collecting this information are provided in Appendix I. These would be modified to be culturally and institutionally relevant for each setting.

Value allocation is a methodology that determines the relative importance of each offering based on the business's willingness to pay for the offering. It grounds the data collection in an actual cost exercise, which brings an element of realism that is absent in other survey questions. We will use a two-step process. The first is to ask if the respondent would purchase the offering. The second then takes positive selections and asks them to allocate a budget across all of those they selected. Example survey questions for collecting this information are provided in Appendix J. These would be modified to be culturally and institutionally relevant for each setting.

We analyze the resulting data to find the aggregate likelihood of purchasing these educational offerings and the relative value of each. These results can then be used to develop a plan for curriculum development that more directly targets the needs of the biotechnology and healthcare technology industries, thus improving the currency and desirability of our curricula for our students and for producing translational research based on academic findings.

V. CONCLUSIONS

The apparent technological knowledge gaps that existed between academia and industry led to the work-in-progress multi-stage curriculum development model to bridge those gaps. We are developing a model that incorporated data analytics with primary research to identify topic areas for curriculum refinement and development. The initial data analytics stages have been productive in elucidating topical needs of industry that can form the milieu for an academic education in informatics, a STEM discipline. Our primary research stage will provide the granularity needed to begin the process of ensuring that our courses meet the needs of multiple stakeholders for our institution. Once validated, this model can be used at all levels in STEM education to ensure the continuing success of our organizations.

REFERENCES.

[1] Kayyali B, Knott D, Kuiken SV. “How big data is shaping US health care,” McKinsey Quarterly (May 2013).

[2] Cattell J, Chilukuri S, Levy M. “How big data can revolutionize pharmaceutical R&D,” *McKinsey & Company* (April 2013).

[3] Brusic V and Ranganathan S. “Critical technologies for bioinformatics”. *Briefings in Bioinformatics*. 2008; 9(4): 261-262.

[4] Niedergassel, Benjamin & Leker, Jens. (2009). “Collaborative R&D Projects: Differences between University-University and University-Industry Partnerships,”

[5] Rybnicek R and Königsgruber R. What makes industry–university collaboration succeed? A systematic review of the literature. *Journal of Business Economics*. 2019; 89: 221-250.

[6] Agrawal, A. (2001) University-to-Industry Knowledge Transfer: Literature Review and Unanswered Questions. International Journal of Management Reviews, 3, 285–302.

[7] Branstetter L, Ogura Y. Is academic science driving a surge in industrial innovation? Evidence from patent citations. *National Bureau of Economic Research*. (2005) 11561. (DOI): 10.3386/w11561.

[8] Mitton C, Adair CE, McKenzie E, Patten SB, Waye PB. Knowledge transfer and exchange: review and synthesis of the literature. *Milbank Quarterly*. 2007; 85(4): 729-68.

[9] Siegel DS, Link AN, Waldman DA, and Atwater LE. Commercial Knowledge Transfers from Universities to Firms: Improving the Effectiveness of University-Industry Collaboration. *The Journal of High Technology Management Research*. 2003; 14(1): 111-133.

[10] Goduscheit, R.C. “How Barriers to Collaboration Prevent Progress in Demand for Knowledge: A Dyadic Study of Small and Medium-Size Firms, Research and Technology Organizations and Universities.” *Creativity and Innovation Management*. 2015; 24:29-54.

[11] Scheitz, C.J.F., Peck L.J., Groan E.S. “Biotechnology software in the digital age: are you winning?” *Journal of Industrial Microbiology & Biotechnology*. 2018; 45: 529-534.

[12] https://www.biocrossroads.com/top100/

[13] Chary M., Parikh S., Manini A.F., Boyer E.W., Radeos M. “A review of natural language processing in Medical Education.” West J Emerg Med. 2019; 20(1):78-86.

[14] Meystre S.M., Savova G.K., Kipper-Schuler K.C.., Hurdle J.F. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearbook Med Inform* 2008; 17(01): 128-144. DOI: 10.1055/s-0038-1638592

[15] McEntire R et.al.. “Application of an automated natural language processing (NLP) workflow to enable federated search of external biomedical content in drug discovery and development.” *Drug Discovery Today*. 2016; 21(5):826-835.

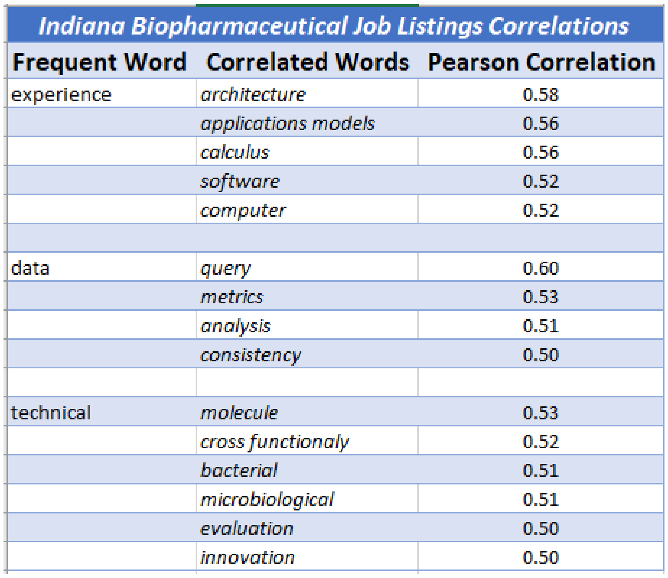
[16] Murff HJ, FitzHenry F, Matheny ME, et al. “Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing.” *JAMA*. 2011;306(8):848–855. doi:10.1001/jama.2011.1204

[17] Szleza´k N, Evers M, Wang J, Perez L.” The role of big data and advanced analytics in drug discovery, development, and commercialization.” *Clin. Pharm. Ther*. 2014; 95: 492–495.

[18] Deyati A, Younesi E, Hoffman-Apitius M, Novac N. (2013) “Challenges and opportunities for oncology biomarker discovery.” *Drug Discovery Today*. 2013; 18: 614-624. DOI: 10.1016/j.drudis.2012.12.011.

[19] Feinerer, I, Hornik, K, Meyer, D. “Text mining infrastructure in R,” Journal of Statistical Software 2008; 25:1-54.

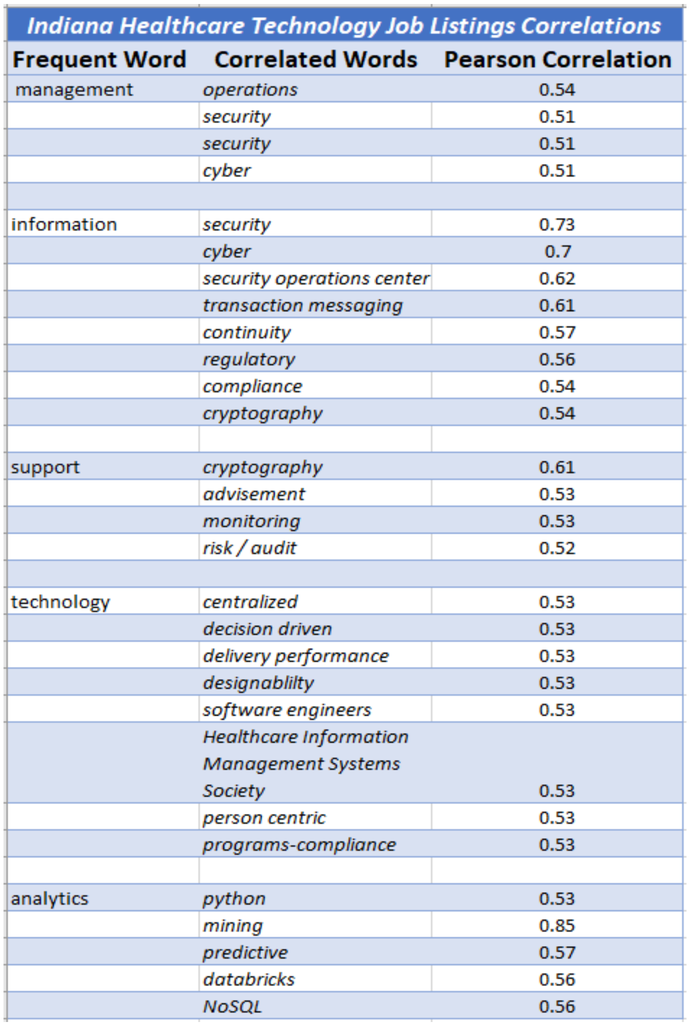
APPENDIX A:



Indiana Biopharmaceutical Job Listings Correlations

Frequent Word	Correlated Words	Pearson Correlation
experience	*architecture*	0.58
	applications models	0.56
	calculus	0.56
	software	0.52
	computer	0.52
data	*query*	0.60
	metrics	0.53
	analysis	0.51
	consistency	0.50
technical	*molecule*	0.53
	cross functionaly	0.52
	bacterial	0.51
	microbiological	0.51
	evaluation	0.50
	innovation	0.50

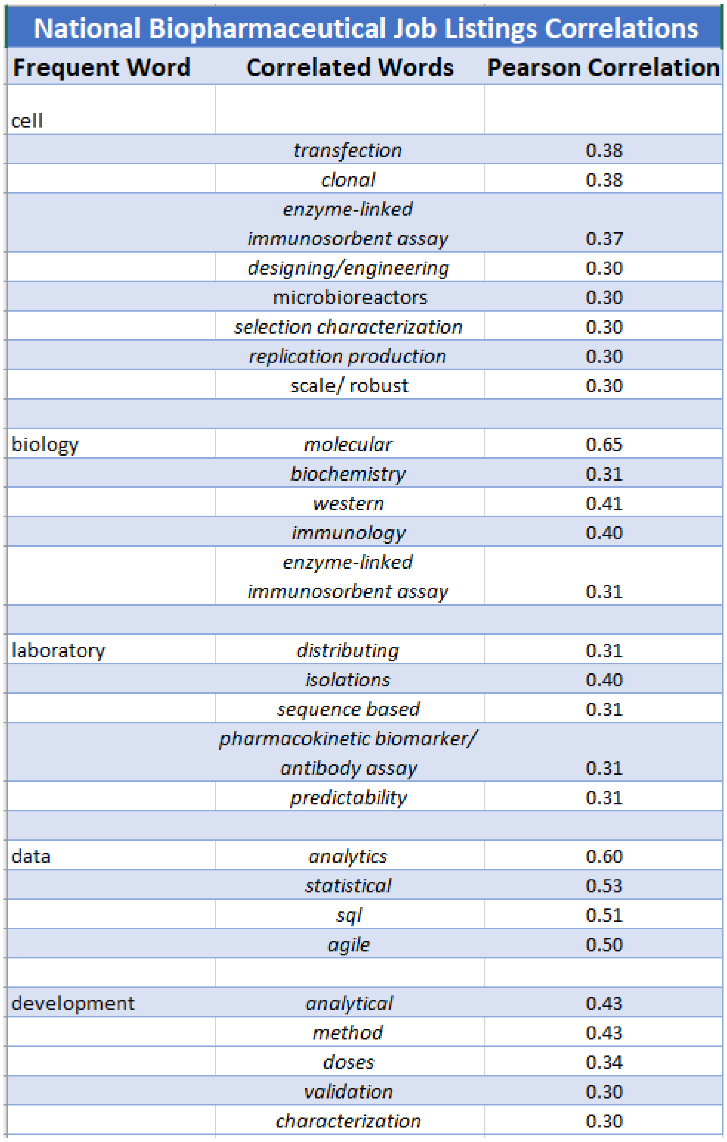
APPENDIX B:



Indiana Healthcare Technology Job Listings Correlations

Frequent Word	Correlated Words	Pearson Correlation
management	*operations*	0.54
	security	0.51
	security	0.51
	cyber	0.51
information	*security*	0.73
	cyber	0.7
	security operations center	0.62
	transaction messaging	0.61
	continuity	0.57
	regulatory	0.56
	compliance	0.54
	cryptography	0.54
support	*cryptography*	0.61
	advisement	0.53
	monitoring	0.53
	risk / audit	0.52
technology	*centralized*	0.53
	decision driven	0.53
	delivery performance	0.53
	designablilty	0.53
	software engineers	0.53
	Healthcare Information Management Systems Society	0.53
	person centric	0.53
	programs-compliance	0.53
analytics	*python*	0.53
	mining	0.85
	predictive	0.57
	databricks	0.56
	NoSQL	0.56

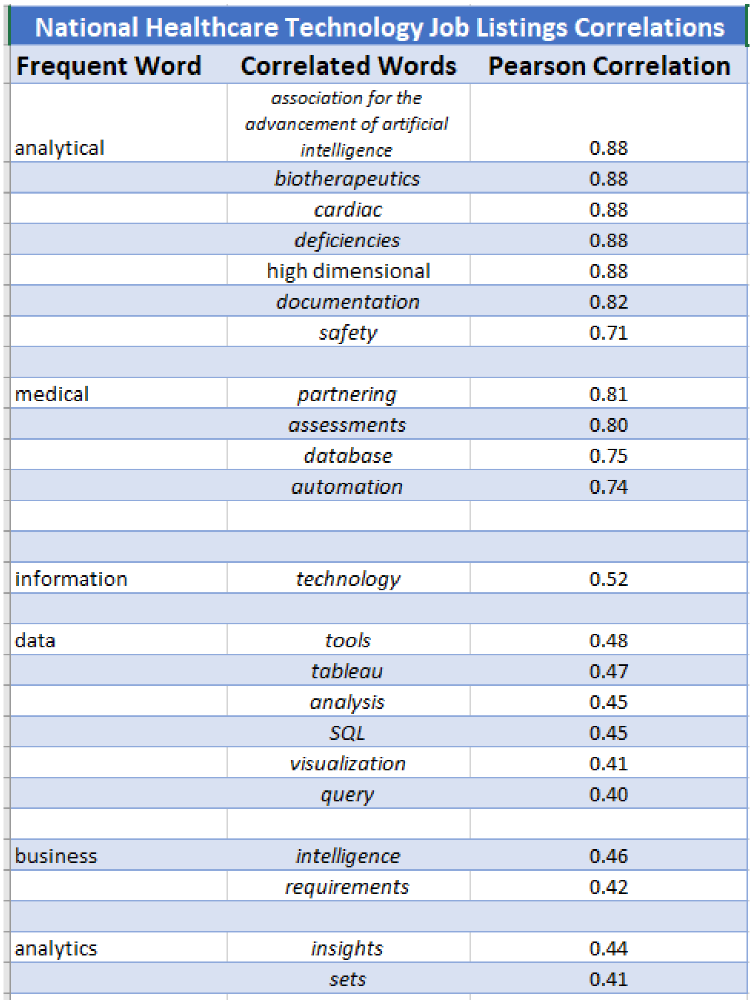
APPENDIX C:



National Biopharmaceutical Job Listings Correlations

Frequent Word	Correlated Words	Pearson Correlation
cell		
	transfection	0.38
	clonal	0.38
	enzyme-linked immunosorbent assay	0.37
	designing/engineering	0.30
	microbioreactors	0.30
	selection characterization	0.30
	replication production	0.30
	scale/ robust	0.30
biology	*molecular*	0.65
	biochemistry	0.31
	western	0.41
	immunology	0.40
	enzyme-linked immunosorbent assay	0.31
laboratory	*distributing*	0.31
	isolations	0.40
	sequence based	0.31
	pharmacokinetic biomarker/ antibody assay	0.31
	predictability	0.31
data	*analytics*	0.60
	statistical	0.53
	sql	0.51
	agile	0.50
development	*analytical*	0.43
	method	0.43
	doses	0.34
	validation	0.30
	characterization	0.30

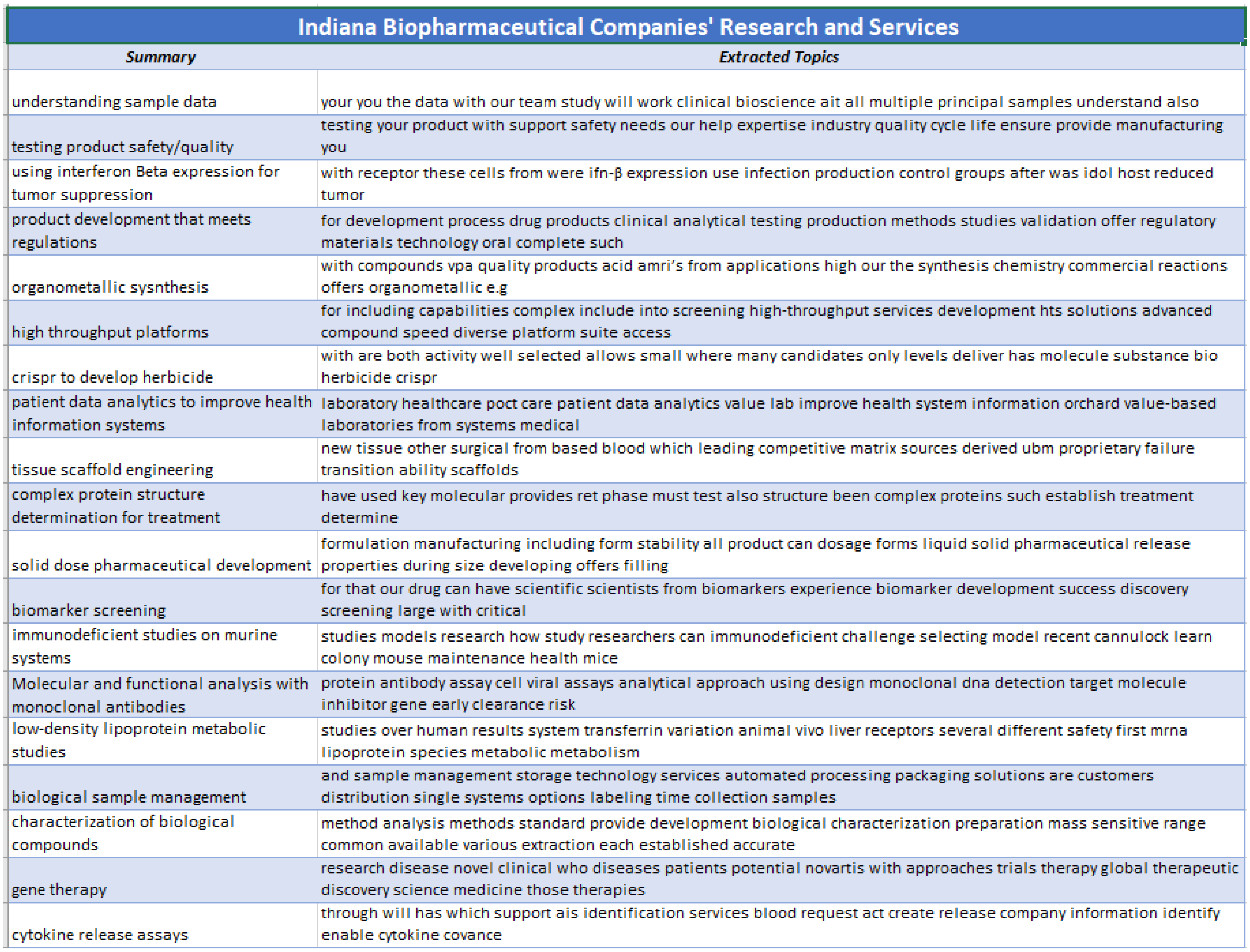
APPENDIX: D



National Healthcare Technology Job Listings Correlations

Frequent Word	Correlated Words	Pearson Correlation
analytical	*association for the advancement of artificial intelligence*	0.88
	biotherapeutics	0.88
	cardiac	0.88
	deficiencies	0.88
	high dimensional	0.88
	documentation	0.82
	safety	0.71
medical	*partnering*	0.81
	assessments	0.80
	database	0.75
	automation	0.74
information	*technology*	0.52
data	*tools*	0.48
	tableau	0.47
	analysis	0.45
	SQL	0.45
	visualization	0.41
	query	0.40
business	*intelligence*	0.46
	requirements	0.42
analytics	*insights*	0.44
	sets	0.41

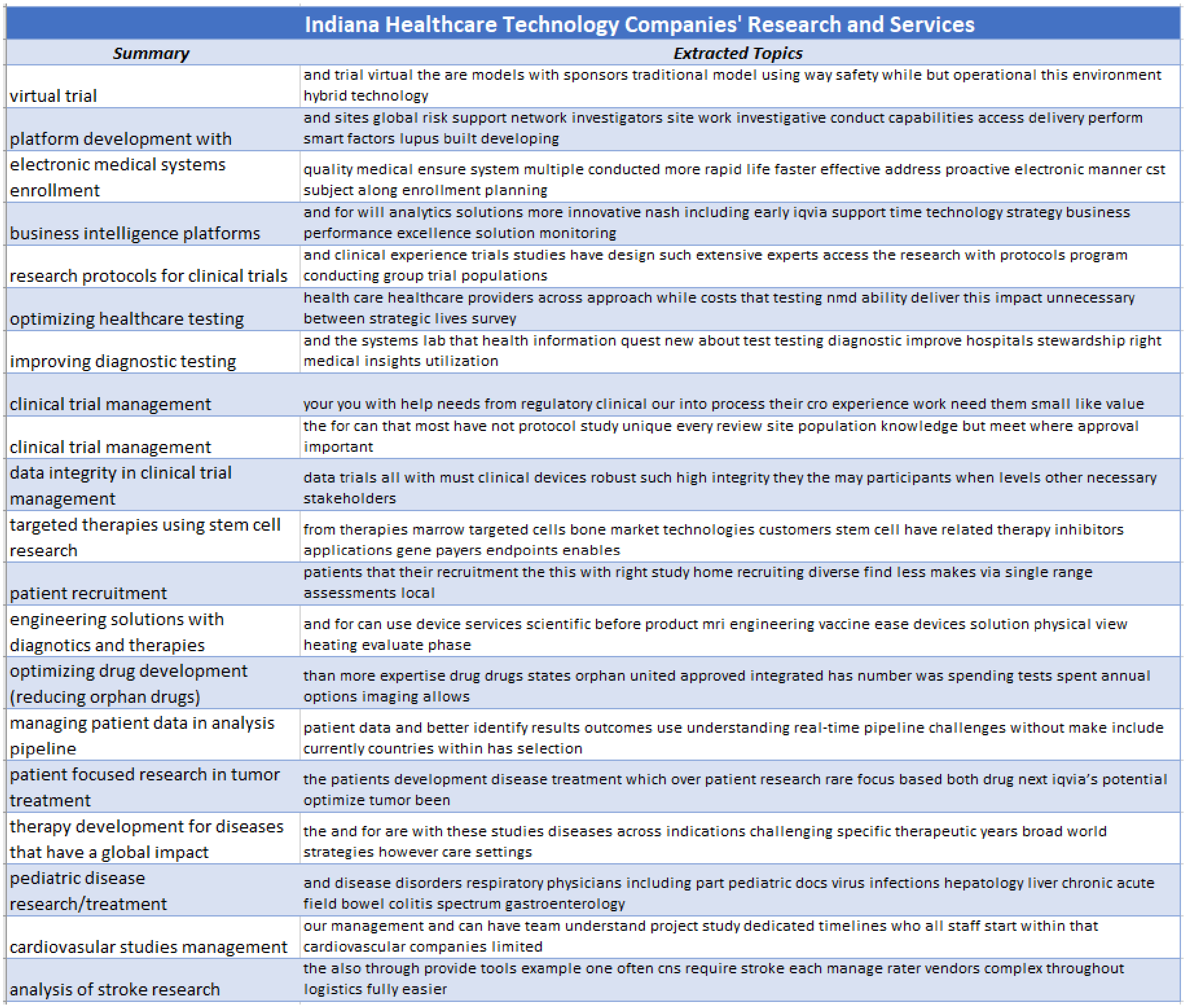
Appendix E:



Indiana Biopharmaceutical Companies' Research and Services

Summary	*Extracted Topics*
understanding sample data	your you the data with our team study will work clinical bioscience ait all multiple principal samples understand also
testing product safety/quality	testing your product with support safety needs our help expertise industry quality cycle life ensure provide manufacturing you
using interferon Beta expression for tumor suppression	with receptor these cells from were ifn-β expression use infection production control groups after was idol host reduced tumor
product development that meets regulations	for development process drug products clinical analytical testing production methods studies validation offer regulatory materials technology oral complete such
organometallic sysnthesis	with compounds vpa quality products acid amri's from applications high our the synthesis chemistry commercial reactions offers organometallic e.g
high throughput platforms	for including capabilities complex include into screening high-throughput services development hts solutions advanced compound speed diverse platform suite access
crispr to develop herbicide	with are both activity well selected allows small where many candidates only levels deliver has molecule substance bio herbicide crispr
patient data analytics to improve health information systems	laboratory healthcare poct care patient data analytics value lab improve health system information orchard value-based laboratories from systems medical
tissue scaffold engineering	new tissue other surgical from based blood which leading competitive matrix sources derived ubm proprietary failure transition ability scaffolds
complex protein structure determination for treatment	have used key molecular provides ret phase must test also structure been complex proteins such establish treatment determine
solid dose pharmaceutical development	formulation manufacturing including form stability all product can dosage forms liquid solid pharmaceutical release properties during size developing offers filling
biomarker screening	for that our drug can have scientific scientists from biomarkers experience biomarker development success discovery screening large with critical
immunodeficient studies on murine systems	studies models research how study researchers can immunodeficient challenge selecting model recent cannulock learn colony mouse maintenance health mice
Molecular and functional analysis with monoclonal antibodies	protein antibody assay cell viral assays analytical approach using design monoclonal dna detection target molecule inhibitor gene early clearance risk
low-density lipoprotein metabolic studies	studies over human results system transferrin variation animal vivo liver receptors several different safety first mrna lipoprotein species metabolic metabolism
biological sample management	and sample management storage technology services automated processing packaging solutions are customers distribution single systems options labeling time collection samples
characterization of biological compounds	method analysis methods standard provide development biological characterization preparation mass sensitive range common available various extraction each established accurate
gene therapy	research disease novel clinical who diseases patients potential novartis with approaches trials therapy global therapeutic discovery science medicine those therapies
cytokine release assays	through will has which support ais identification services blood request act create release company information identify enable cytokine covance

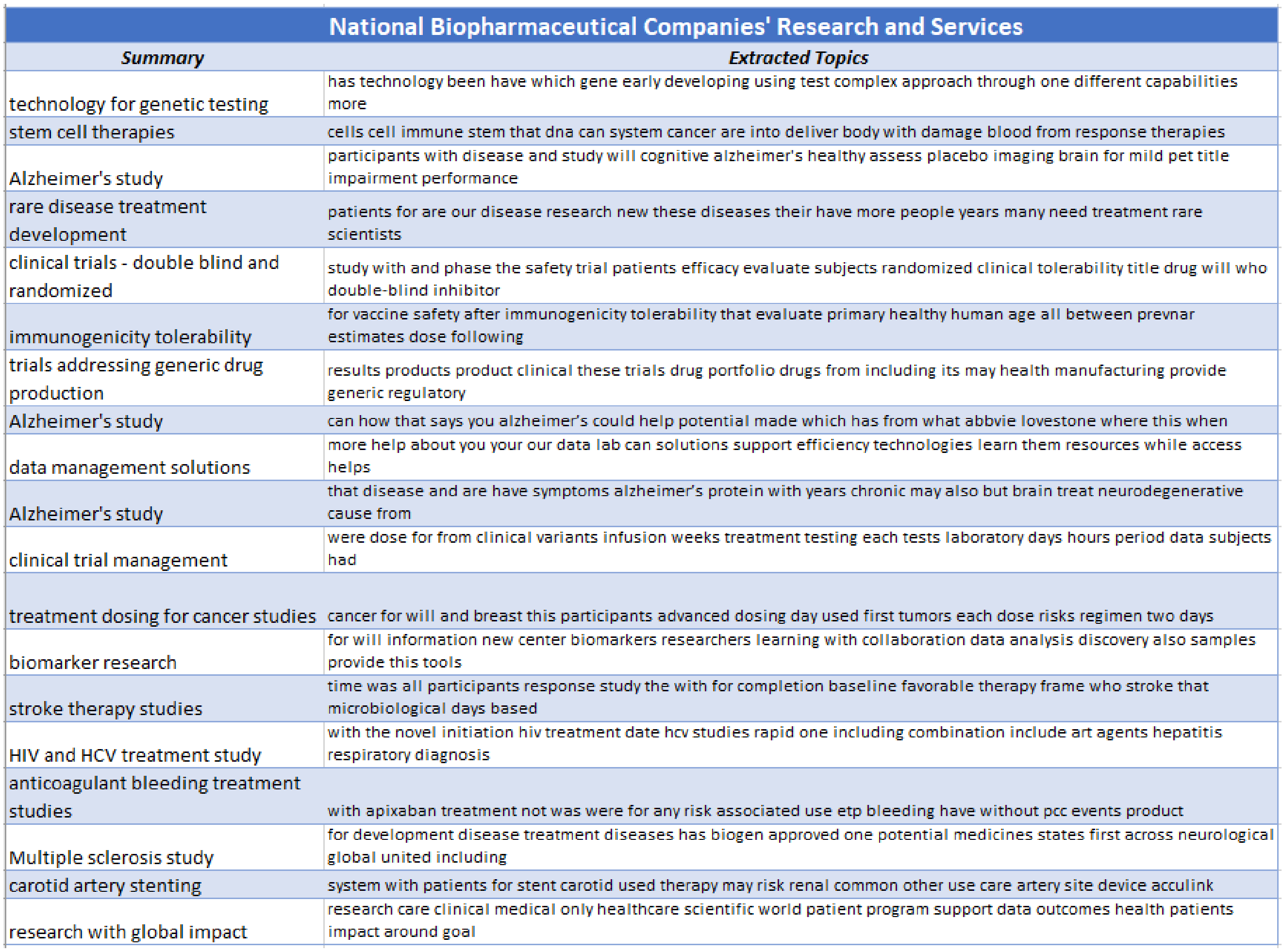
APPENDIX F:



Indiana Healthcare Technology Companies' Research and Services

Summary	*Extracted Topics*
virtual trial	and trial virtual the are models with sponsors traditional model using way safety while but operational this environment hybrid technology
platform development with	and sites global risk support network investigators site work investigative conduct capabilities access delivery perform smart factors lupus built developing
electronic medical systems enrollment	quality medical ensure system multiple conducted more rapid life faster effective address proactive electronic manner cst subject along enrollment planning
business intelligence platforms	and for will analytics solutions more innovative nash including early iqvia support time technology strategy business performance excellence solution monitoring
research protocols for clinical trials	and clinical experience trials studies have design such extensive experts access the research with protocols program conducting group trial populations
optimizing healthcare testing	health care healthcare providers across approach while costs that testing nmd ability deliver this impact unnecessary between strategic lives survey
improving diagnostic testing	and the systems lab that health information quest new about test testing diagnostic improve hospitals stewardship right medical insights utilization
clinical trial management	your you with help needs from regulatory clinical our into process their cro experience work need them small like value
clinical trial management	the for can that most have not protocol study unique every review site population knowledge but meet where approval important
data integrity in clinical trial management	data trials all with must clinical devices robust such high integrity they the may participants when levels other necessary stakeholders
targeted therapies using stem cell research	from therapies marrow targeted cells bone market technologies customers stem cell have related therapy inhibitors applications gene payers endpoints enables
patient recruitment	patients that their recruitment the this with right study home recruiting diverse find less makes via single range assessments local
engineering solutions with diagnotics and therapies	and for can use device services scientific before product mri engineering vaccine ease devices solution physical view heating evaluate phase
optimizing drug development (reducing orphan drugs)	than more expertise drug drugs states orphan united approved integrated has number was spending tests spent annual options imaging allows
managing patient data in analysis pipeline	patient data and better identify results outcomes use understanding real-time pipeline challenges without make include currently countries within has selection
patient focused research in tumor treatment	the patients development disease treatment which over patient research rare focus based both drug next iqvia's potential optimize tumor been
therapy development for diseases that have a global impact	the and for are with these studies diseases across indications challenging specific therapeutic years broad world strategies however care settings
pediatric disease research/treatment	and disease disorders respiratory physicians including part pediatric docs virus infections hepatology liver chronic acute field bowel colitis spectrum gastroenterology
cardiovasular studies management	our management and can have team understand project study dedicated timelines who all staff start within that cardiovascular companies limited
analysis of stroke research	the also through provide tools example one often cns require stroke each manage rater vendors complex throughout logistics fully easier

APPENDIX G:



National Biopharmaceutical Companies' Research and Services	
Summary	***Extracted Topics***
technology for genetic testing	has technology been have which gene early developing using test complex approach through one different capabilities more
stem cell therapies	cells cell immune stem that dna can system cancer are into deliver body with damage blood from response therapies
Alzheimer's study	participants with disease and study will cognitive alzheimer's healthy assess placebo imaging brain for mild pet title impairment performance
rare disease treatment development	patients for are our disease research new these diseases their have more people years many need treatment rare scientists
clinical trials - double blind and randomized	study with and phase the safety trial patients efficacy evaluate subjects randomized clinical tolerability title drug will who double-blind inhibitor
immunogenicity tolerability	for vaccine safety after immunogenicity tolerability that evaluate primary healthy human age all between prevnar estimates dose following
trials addressing generic drug production	results products product clinical these trials drug portfolio drugs from including its may health manufacturing provide generic regulatory
Alzheimer's study	can how that says you alzheimer's could help potential made which has from what abbvie lovestone where this when
data management solutions	more help about you your our data lab can solutions support efficiency technologies learn them resources while access helps
Alzheimer's study	that disease and are have symptoms alzheimer's protein with years chronic may also but brain treat neurodegenerative cause from
clinical trial management	were dose for from clinical variants infusion weeks treatment testing each tests laboratory days hours period data subjects had
treatment dosing for cancer studies	cancer for will and breast this participants advanced dosing day used first tumors each dose risks regimen two days
biomarker research	for will information new center biomarkers researchers learning with collaboration data analysis discovery also samples provide this tools
stroke therapy studies	time was all participants response study the with for completion baseline favorable therapy frame who stroke that microbiological days based
HIV and HCV treatment study	with the novel initiation hiv treatment date hcv studies rapid one including combination include art agents hepatitis respiratory diagnosis
anticoagulant bleeding treatment studies	with apixaban treatment not was were for any risk associated use etp bleeding have without pcc events product
Multiple sclerosis study	for development disease treatment diseases has biogen approved one potential medicines states first across neurological global united including
carotid artery stenting	system with patients for stent carotid used therapy may risk renal common other use care artery site device acculink
research with global impact	research care clinical medical only healthcare scientific world patient program support data outcomes health patients impact around goal

APPENDIX H:

National Healthcare Technology Companies' Research and Services	
Summary	***Extracted Topics***
N/A	that are more have been about than has for use their better because there companies could often its well make
toxicity study for oncological therapy	car-t therapy i-o therapies toxicity treatment these all cost approval oncologists r/r first fda significant adoption might will effects stakeholders
immuotherapy in tumor treatment	patients with had response are tumor tmb immunotherapy more respond treated who survival approach while benefit complete immune overall
cardiac surgery mortality study	for has new when surgery gulf hospital rates reduce cardiovascular mortality assessment death screening massachusetts risk significantly innovative
cardiovascular study recruitment	data their how care participants would that our collect value heart team for stress out research rate value-based
oncology research in patient care	for patients health research care this oncology pro all our medical practices practice community studies physicians network support cardinal
insulin involvement in prostate cancer	cancer with that drugs her drug prostate tumor has colon levels used insulin cell inhibitor protein combination growth signaling
analysis of adenomas treatments	were was with and than among years months using lower more higher after model range associated adr quality odds
coronavirus infection study	with reported for this have infection cases from states several been respiratory china other similar covid days year
leukemia study	clinical treatment pathways for were should trials process trial identified committee leukemia only most patient pathway testing therapy event
N/A	research cancer national medicine published institute phd clinical health professor ph.d hopkins johns center department university oncology foundation
colibactin study	that cells was cell which then from into found also could like colibactin protein cellular using lab receptor antigen
cardiac treatment study	with and study this treatment patients heart disease group population device events whether failure there for associated will cardiac
optimizing patient experience via technology	with for design their experience disease you patient this are healthcare people into technology process products chronic like needs
modeling human diseases in mice	are human how that can blood new they system diseases mice cells all not even what other models
research into genetic resilience	from for with was study risk genetic resilience not significant found therapeutic family potential one research association known
data collection optimization	data healthcare real-world can value from outcomes care have for collect need collection improve ability also requires
tumor cell sequencing analysis in pediatric cancer	cancer tumor for sequencing tumors pediatric mutations jude solid patients genome analysis are said cancers researchers different
N/A	can will may they but these help which have some people not their also time most
research into brain disorders	are such but understanding may not important disease which factors severe critical treat risk brain guidelines acute major

APPENDIX I:

Example Survey Items for conducting Gap Analysis

1. How important is it to you that your employees have the ability to…

	Not at all important						Extremely Important
	1	2	3	4	5	6	7
Informatics Skill 1	❍	❍	❍	❍	❍	❍	❍
Informatics Skill 2	❍	❍	❍	❍	❍	❍	❍
Informatics Skill N	❍	❍	❍	❍	❍	❍	❍

2. Now, how satisfied are you with the educational opportunities available for Indiana professionals in…

	Not at all Satisfied							Extremely Satisfied
	1		2	3	4	5	6	7
Informatics Skill 1	❍		❍	❍	❍	❍	❍	❍
Informatics Skill 2	❍		❍	❍	❍	❍	❍	❍
Informatics Skill N	❍		❍	❍	❍	❍	❍	❍

APPENDIX J:

Example Survey Items for conducting Value Allocation

1. BHI is thinking of developing the following educational modules in informatics for the local community in the health/bioscience industries. Please read and think about this these proposed offerings and let us know the following…

	Would you engage with DBHI to purchase this training?
NEW TRAINING PROGRAMS	
a. Informatics Offering 1	❑YES ❑ NO
b. Informatics Offering 2	❑YES ❑ NO
c. Informatics Offering N	❑YES ❑ NO

2. Now let’s revisit the offerings that you said you would purchase from DBHI. Assume that your educational budget was $100. How would you allocate your budget across these programs based on how valuable it would be to your business or organization?

How much would you spend out of $100	Dollars
Informatics Offering 1	
Informatics Offering 2	
Informatics Offering N	