[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

J Pathol Inform

Editor-in-Chief:
Anil V. Parwani , Pittsburgh, PA, USA
Liron Pantanowitz, Pittsburgh, PA, USA
OPEN ACCESS
HTML format
For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Editorial

How can we improve Science, Technology, Engineering, and Math education to encourage careers in Biomedical and Pathology Informatics?

Rahul Uppal[1], Gunasheil Mandava[1], Katrina M. Romagnoli[1], Andrew J. King[1], Amie J. Draper[1], Adam L. Handen[1], Arielle M. Fisher[1], Michael J. Becich[1], Joyeeta Dutta-Moscato[1]

Department of Biomedical Informatics, University of Pittsburgh, School of Medicine, Pittsburgh, USA

E-mail: *Joyeeta Dutta-Moscato - jod30@pitt.edu
*Corresponding author

Received: 22 July 2015 Accepted: 16 October 2015 Published: 29 January 2016

Abstract

The Computer Science, Biology, and Biomedical Informatics (CoSBBI) program was initiated in 2011 to expose the critical role of informatics in biomedicine to talented high school students.[1] By involving them in Science, Technology, Engineering, and Math (STEM) training at the high school level and providing mentorship and research opportunities throughout the formative years of their education, CoSBBI creates a research infrastructure designed to develop young informaticians. Our central premise is that the trajectory necessary to be an expert in the emerging fields of biomedical informatics and pathology informatics requires accelerated learning at an early age.In our 4th year of CoSBBI as a part of the University of Pittsburgh Cancer Institute (UPCI) Academy (http://www.upci.upmc.edu/summeracademy/), and our 2nd year of CoSBBI as an independent informatics-based academy, we enhanced our classroom curriculum, added hands-on computer science instruction, and expanded research projects to include clinical informatics. We also conducted a qualitative evaluation of the program to identify areas that need improvement in order to achieve our goal of creating a pipeline of exceptionally well-trained applicants for both the disciplines of pathology informatics and biomedical informatics in the era of big data and personalized medicine.

Key words: Bioinformatics, Computer Science, Biology, and Biomedical Informatics, pathology informatics, Science, Technology, Engineering, and Math

Access this article online
Website:
www.jpathinformatics.org
DOI: 10.4103/2153-3539.175375
Quick Response Code:

INTRODUCTION

"Big Data," a popular buzzword dating back to 2001, has rendered a need for savvy information scientists and machine learning experts. IBM states that new skills are needed to fully harness the power of big data, and while courses are being offered to train a new generation of data experts, it will take time to build them into the workforce.[2] In an effort to expedite this process, CoSBBI aims to introduce high school students to the

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

This article may be cited as: Uppal R, Mandava G, Romagnoli KM, King AJ, Draper AJ, Handen AL, *et al.* How can we improve Science, Technology, Engineering, and Math education to encourage careers in Biomedical and Pathology Informatics?. J Pathol Inform 2016;7:2.
Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2016/7/1/2/175375

© 2015 Journal of Pathology Informatics | Published by Wolters Kluwer - Medknow

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

practice of informatics in medicine, including the use of computational techniques to solve biological problems. We believe that CoSBBI, and other similar programs, can be used to select and train this new generation of data scientists, thus preparing them for the workforce with the tools, experience, and professional network required to succeed in the domain.

We reported on our 2013 CoSBBI experience[1] describing the program's mission and curriculum. In the CoSBBI class of 2013, 11 scholars developed and presented projects spanning a broad range of topics including bioinformatics, pathology informatics, computational biology, machine learning, image analysis, pharmacogenomics, and telemedicine. These scholars and their faculty research mentors were encouraged to publish their abstracts in the Journal of Pathology Informatics (JPI).[3] Some of those students are now attending college at the California Institute of Technology, Carnegie Mellon University, University of Notre Dame, University of California at Los Angeles, and our very own University of Pittsburgh. Further, our program graduates continue to demonstrate success in informatics an example of which is an alumnus that won the "Pitt Smash Mash!" student start-up competition for an app that serves as a conduit between students and the University Health services. Another alumnus completed a summer research internship at Duke University and was selected from a competitive pool of applicants to present her research at the American Medical Informatics Association (AMIA) 2014 annual symposium (http://www.amia.org/amia2014/high-school-scholars).

This year, we continued our mission of introducing biomedical informatics through a STEM-oriented research academy. We began with a statement of our goal: To provide young talent with a survey of fundamentals, exposure to current informatics research, and a research internship experience. As we continue to build on these goals every year, we have enhanced the 2014 CoSBBI curriculum to meet the evolving needs of our students; this editorial is a synopsis of these changes, including a qualitative evaluation of the 2014 CoSBBI program.

CLASSROOM INSTRUCTION

As in last year's CoSBBI program, the classroom portion was designed to provide a didactic introduction to biomedical informatics, promote an understanding of research, and expose scholars to career opportunities in the field. The complete syllabus and teaching materials for the 8 weeks CoSBBI program can be accessed at: http://faculty.dbmi.pitt.edu/cosbbi/cosbbi2014/. Once again, we used the online, open-access Translational Bioinformatics compilation (PLOS Computational Biology Collection, www.ploscollections.org/translationalbioinformatics) as the primary textbook. In the first few days, scholars attended orientation sessions and were given guest passes to attend the 2014 National Library of Medicine Informatics Training Conference, which was hosted by the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh. For the remainder of the first 6 weeks, the scholars spent two hours each day in the classroom learning about the fundamentals of informatics in a variety of domains. To better equip students with the necessary skills for completing an independent informatics research project, we implemented a week-long, hands-on programing boot camp. It was conducted during the 1st week of didactic sessions and was a significant addition to the 2014 curriculum. Following the boot camp, each classroom session was comprised of one instructional hour led by doctoral, postdoctoral, and medical fellows, and 1 h of research presentation and discussion led by academic researchers and industry guests. Lectures in the early weeks covered basics of molecular biology, bioinformatics tools, computational thinking, statistics, and data mining, while lectures in later weeks focused on specific areas of biomedical informatics.

The programing boot camp was designed to provide students with a brief introduction to programing and its various applications within the field of biomedical informatics. More specifically, the course was designed to help students: (1) understand basic programing concepts and how to implement those concepts, (2) recognize some basic programing solutions for real-world applications in biomedical informatics, and (3) explore additional languages and more advanced programing topics with a higher level of comfort. Specific topics covered in the programing boot camp included data types, Boolean logic, loops, data structures, file input/output, functions, and code libraries for bioinformatics. Students were instructed in the Python programing language, which is commonly used in biomedical informatics and other disciplines. The instruction was highly interactive, with short lectures interspersed with demonstrations during the 1st h, followed by working through programing problem sets in small, assisted groups during the 2nd h. Problem sets were related to lecture material and covered real-world problems in biomedical informatics (e.g., gene detection, elementary clinical decision support). CodeAcademy (http://www.codecademy.com/) was used to implement the problem sets, as it offers basic Python tutorials, is available online, and requires no installation. In addition, CodeAcademy allowed us to design our own problem sets and create detailed tutorials and error messages to help students work through the assignments successfully.

INTRODUCING SCHOLARS TO RESEARCH

While the classroom sessions were focused on concepts and application, the majority of the scholar's time was

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

reserved for pursuing deeper skills relevant to their individual research project. Upon acceptance into CoSBBI, each scholar was matched to a faculty mentor involved in informatics research at the University of Pittsburgh. The mentor/mentee matches are based on the scholars' background and interests as stated in their application and the availability of suitable mentors. Every attempt is made to making this matching process synergistic with the scholars stated career goals. We wanted to expose the CoSBBI scholars to ongoing, hands-on scientific research in Biomedical and Pathology Informatics. In addition to the classroom sessions, we wanted to focus on the development of three primary areas of research skills: (1) reading, evaluation, and presentation of current literature; (2) conducting independent research in a timely manner; and (3) communication of research through scientific writing and presentation. Our approach to these three primary areas of development is discussed below.

We felt that it was important to demonstrate the importance of reviewing current literature by teaching the students the necessary skills to read, critically evaluate, and present peer-reviewed papers. Toward this end, scholars selected and presented a peer-reviewed article. In this journal club style session, they were encouraged to select scholarly papers relevant to their specific research question with the help of their faculty mentor.

Midway through the program, weekly meetings began to provide students with a forum to discuss the progress of their individual research and receive feedback from their peers and members of the DBMI. This provided the students with milestones toward completion of their project, along with close mentorship and peer evaluation required to complete a sophisticated research project on time.

Finally, we used several classroom sections to train the students how to communicate their research findings through scientific writing and presentation. Near the end of the program, the scholars applied these skills to write scientific abstracts summarizing their work, which can be found at the end of this editorial. On the final day of the program, the scholars gave oral presentations to an open audience at DBMI and presented posters at the UPCI Academy closing symposium.

PROGRAM EVALUATION

Two summer interns (one high school student [GM] and one college student [RU]) developed a survey instrument [Supplement 1] with two doctoral students (KMR, JD-M), and conducted interviews with all 2014 CoSBBI students ($n = 9$). Interviews were recorded, transcribed, and analyzed thematically to identify major successes and problems with the program. We discuss here important observations that emerged as general themes not only relevant for our own improvement, but also as considerations for any Biomedical or Pathology Informatics Department interested in STEM level outreach.

Overall, students viewed the program favorably, rating CoSBBI an average of 4.1 out of 5. They said they were glad that they had received exposure to the field of biomedical informatics (specifically: career paths, research, and work life as a researcher) while learning key skills such as time management, teamwork, and self-discipline. Engaging lectures involving hands-on demonstrations and activities were considered to be the most interesting. The scholars rated their research projects (3.7 out of 5) and research mentors (4.1 out of 5) favorably, with individual comments of dissatisfaction ranging in the areas of communication, compatibility, choice in project formulation, and feeling overwhelmed or unchallenged. The majority of scholars reported feeling that their project was interesting and engaging, and they were able to see its applicability in the domain of informatics. While the sources of dissatisfaction cited by scholars were quite context-specific and mostly relevant to individual scholar-mentor partnerships, one common remark was that their projects did not pertain to bioinformatics. One student, whose project was in clinical informatics said, "My mentor told me on the 1st day 'I don't do bioinformatics, I have nothing related to that; so that's where the problem lies,' that my mentor didn't really deal with bioinformatics as his main study." This revealed an interesting contradiction we had been unknowingly propagating to the scholars: as presented in our editorial last year[1] we promote undergraduate programs in bioinformatics as a route to a future career in biomedical informatics. However, as most practitioners of biomedical informatics know, bioinformatics is not necessarily the ideal foundation for careers in all areas of biomedical informatics. We delve into this further as we discuss our thoughts toward finessing the pipeline.

Among areas of improvement recommended for our program, one of the recurring themes was a desire for more structure in their daily schedule. The students at other sites at the UPCI summer academy are mostly assigned to laboratory-based projects which involve bench work. In the DBMI, however, projects mostly require solitary work on a computer. We had thought of limiting their classroom time in favor of leaving them more time to focus on their individual research projects; many of our respondents perceived this as "too much free time." It is interesting to note that many respondents also felt they were not making adequate research progress until late in the program.

Being predominantly involved in graduate level education, we may have been too reliant on a scholar's

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

personal initiative in their individual projects, particularly compared to the demanding schedules that high performing adolescents are used to. One idea to address this issue is to incorporate a schedule of research reporting in the classroom from the very beginning of the program. This will encourage peer mentoring and involvement from early stages of the research project. Another idea is to have mentoring teams for every student. Integrating a multilevel team of postdoctoral fellows and/or advanced graduate students in the faculty mentor's laboratory to be daily comrades and mentors to the scholars will allow closer monitoring in addition to providing a more immersive educational experience.

A common theme in the feedback about classroom instruction also implied an expectation of more structured guidance. In keeping with our efforts to let scholars prioritize research work, we did not have homework or exams on a regular schedule. Problem sets were assigned only for the programing and statistics lectures. Readings were assigned for every topic, but students were not specifically tested on them. From our perspective, this was to be the first step into the world of independence in scholarly pursuit. We purposely allowed scholars to choose areas that they wished to focus more time on, while providing them with exposure at a broader level, as well as direction on where to go for more. Some scholars thrived with this freedom, but many others perceived this as a lack of rigor, wishing for more of a challenge in the classroom. To address this issue, we will consider administering weekly quizzes. It would be important, though, to craft tests that are meaningful toward a cohesive preinformatics foundation. Accordingly, the testing emphasis should be on broader concepts, not pedantry.

FINESSING THE PIPELINE

In service to our priority of contributing to a pipeline for biomedical informatics, we previously[1] highlighted the emergence of undergraduate Bioinformatics Departments. We encourage CoSBBI students to explore bioinformatics as a college major, but we are also aware that there are alternative majors of foundational value to the study of biomedical informatics that may appeal more broadly to students' career interests. An eloquent opinion piece by Dr. William Hersh discusses this quandary faced in advising students on the most appropriate "preinformatics" college major (http://informaticsprofessor.blogspot.com/search?q=pre-informatics).

Our thoughts on an optimal solution are two-fold: We encourage our students to use their summer at CoSBBI to find what inspires them, and we offer paid summer internships at DBMI for any of them who contact us in following years to pursue further research in biomedical or pathology informatics. Two of last year's students returned as interns, one of whom is majoring in computer science, while the other pursues a premed major with coursework in design and business. Studies have shown that increasing engagement and interest in STEM fields during the precollege years is most effective in cultivating graduates in these fields, even more so than high grades and enrollment in advanced level classes.[4]

Critical to CoSBBI's continued success in encouraging careers in Biomedical and Pathology Informatics is continued mentorship from faculty in both of these disciplines. The DBMI (http://www.dbmi.pitt.edu) and the Division of Pathology Informatics and the Center for Pathology Informatics (http://path.upmc.edu/cpi/) are committed to long-term mentoring of CoSBBI scholars. To date, Dr. Becich has written over 30 letters of recommendation to CoSBBI students and dozens of other letters have been supplied by CoSBBI scholar-mentors. Mentors of CoSBBI scholars have provided ongoing input to college major selection encouraging students to enroll in programs in bioinformatics.[1] In 2013, there were 34 such programs in US colleges and Universities,[1] currently, there are over 60 such programs. Another important area of continued mentorship is the CoSBBI internship program which will be the subject of another JPI article soon to be submitted. The CoSBBI high school scholars are guaranteed paid internships in the DBMI once they successfully complete the 8 weeks summer program. To date, of the nearly 30 students that have participated in CoSBBI, eleven have returned to do paid internships. This internship program will be expanded to include the CoSBBI Innovation Internship (Becich, Becich and Boone, manuscript submitted) which focuses on academic, commercial entrepreneurship. In short, CoSBBI is a unique mentorship program which aims to "pipeline" highly trained students for careers in Bioinformatics, Biomedical, and Pathology Informatics.

CONCLUSIONS

Young scholars exit the CoSBBI program with an exceptional first-hand STEM experience. The experience was enhanced by this year's addition of a computer programing boot camp. The boot camp helped to kick start many of the scholar's research projects and had benefits for both proficient and first-time programers alike. We will continue to fine-tune the curriculum and other aspects of the schedule to best benefit the needs and desires of each scholar. We remain committed to providing the best and brightest high school students the opportunity to find their passion in Biomedical and Pathology Informatics. It is our intention for the program to remain free so that students of all backgrounds may continue to participate. In future years, we hope to have more of our scholars compete in the newly created high school student competition at the AMIA annual

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

symposium and hope for more of them to continue their research projects into their undergraduate years. There is no single solution for inspiring the next generation of scientists, but programs, such as CoSBBI, that introduce high school students to the complex world of STEM are vital. The best evidence of the success of CoSBBI is that over 30% of our students return to do paid internships. The addition of the CoSBBI Innovation Internship will certainly increase the interest in STEM through a focus on commercial entrepreneurship.

Acknowledgments

We acknowledge David Boone, Ph.D. for helpful suggestions and comments, Megan Houlihan for CoSBBI program coordination with the main UPCI Academy, Dr. Michael T. Lotze for his leadership and devotion to the UPCI Academy, Victoria Khersonsky for her technical support for CoSBBI scholars, DBMI graduate students and faculty/staff mentors.

Financial Support and Sponsorship

This CoSBBI track of the UPCI Academy is supported by NIH grants R01 LM010950 from the National Library of Medicine, R01 GM100387 from the National Institute of General Medical Sciences, Doris Duke Foundation, the National Cancer Institute CURE Program (3P30CA047904-22S1) and support from the Jack Kent Cook Foundation. In addition, this program would not be possible without the infrastructure and support teams from the DBMI NLM Training Program Grant in Biomedical Informatics (T15 LM007059), the University of Pittsburgh Cancer Institute (UPCI) Cancer Center Support Grant for the Cancer Bioinformatics Service (P30 CA47904), the Clinical and Translational Science Institute Biomedical Informatics Core (UL1 RR024153), as well as from funds and in-kind services from UPCI and DBMI, and the Departments of Surgery, Immunology, Computational and Systems Biology, Gynecology, and Pharmacology.

Conflicts of Interest

There are no conflicts of interest.

REFERENCES

1. Dutta-Moscato J, Gopalakrishnan V, Lotze MT, Becich MJ. Creating a pipeline of talent for informatics: STEM initiative for high school students in computer science, biology, and biomedical informatics. J Pathol Inform 2014;5:12.
2. Available from: http://www.ibm.com/big-data/us/en/. [Last accessed on 2015 Jan 31].
3. Available from: http://www.jpathinformatics.org. [Last accessed on 2015 Jan 31].
4. Maltese AV, Tai RH. Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. students. Sci Educ 2011;95:877-907.

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

ABSTRACTS

Can varying the salutation and subject lines of E-mail prompts increase patient log-ins to an internet support group for mood and anxiety disorders?

Nikhil R. Cherukupalli[1], Akash Bansal[1], Bea Herbeck-Belnap[2,3], Christopher Wiltrout[2], Bruce L. Rollman[1,2,3,4]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Medicine, Center for Research on Health Care, [3]Division of General Internal Medicine, University of Pittsburgh, [4]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

E-mail: *Bruce L. Rollman - RollmanBL@upmc.edu
*Corresponding author

Context: Internet support groups (ISGs) can enable member-patients to exchange information and emotional support and are a widely-available self-help resource. Yet, sustaining patient engagement on ISGs is challenging. We examined the impact of varying the salutation and subject lines of E-mail prompts on member log-ins to an ISG created as part of an NIMH-funded trial for treating mood and anxiety disorders in primary care.

Technology: The Trial ISG was created in WordPress and hosted on a University server. We created our test messages in Microsoft Outlook and sent them to recipient members via blind mail-merge E-mails to preserve patient confidentiality.

Design: Since 8/1/12, we randomized protocol-eligible depressed and anxious patients recruited from 26 UPMC-affiliated primary care practices have been randomized to one of three groups, including one with password access to our ISG. We analyzed server logs to identify member-patients who had not logged-in to the ISG within the 6 weeks and divided them alphabetically into one of four groups (2 × 2 design). Two groups received: (a) personally addressed (e.g. "Dear John") or generic E-mail messages; and (b) command ("Log in Now!") or collaboratively worded subject lines ("Join the Conversation!"). We sent these messages 6 times over a 2½ weeks period (7/21–8/6), and then examined our server logs to assess their impact on logins (8/7/14).

Results: Of the 280 patients randomized to the ISG as of 7/15/14, 204 (73%) never logged-in after 6/1/14. Of these 204 ISG members who received our E-mail prompts, just 2.5% (5/204) logged into the site: 4% (2/51) from each the "collaborative-personalized" and "commanding-generic" groups and 2% (1/51) from the "commanding-personalized" group.

Conclusions: Increasing ISG member engagement is challenging as both our experimental salutations and subject line variations had essentially no impact on ISG log-ins among those who had not recently logged-in. Nevertheless, we will continue to iterate our E-mail prompt strategy to identify effective strategies at increasing member log-ins and subsequent engagement (e.g., varying different colors, graphics, and timing).

Automated image analysis for immunohistochemical evaluation of protein expression levels to assess their use as biomarkers for prostate cancer

Sahil Dadoo[1], Marianne Notaro[2], Tony Green[2], Anil V. Parwani[2]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Pathology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

E-mail: *Anil V. Parwani - parwaniav@upmc.edu
*Corresponding author

Context: Prostate Cancer is the second most common form of cancer among men, only behind skin cancer. Historically, African-American men have much higher rates, and more aggressive forms, of prostate cancer than other races, but the cause remains unknown. The goal of this project was to evaluate expression levels of two genes to assess their usefulness as biomarkers for prostate cancer.

Technology: Automated image analysis was conducted using Aperio ImageScope, Leica Biosystems. All tissue microarrays were scanned using an aperio scanscope scanner.

Design: 49 cases of African-American prostate cancer patients were selected for the study. Tissue samples from each case were selected to create a tissue microarray for differential analysis. Clinical and pathological information for each tissue sample were annotated in a database. Two antibodies, p38 and STAT3, were analyzed on the tissue microarray for their expression levels using immunohistochemistry. The analysis was completed using automated image analysis software.

Results: STAT3 antibodies produced significantly larger positivity values than p38 antibodies, with an average difference of 0.3351. STAT3 also showed differences in

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

positivity values between the Gleason scores and the five tumor stages. STAT3 portrayed higher positivity values for the less aggressive Gleason scores (0.6178 and 0.5873), and significantly lower positivity values for the more aggressive Gleason scores (0.3968 and 0.4444). For p38, our studies revealed that p38 immunostaining did not significantly differentiate between Gleason grades and tumor stage.

Conclusions: STAT3 antibody may play an important role in prostate neoplasia and, based on the data highlighted in this study, it may play an important role as a novel diagnostic and/or diagnostic biomarker. For p38, additional testing with a larger patient population, as well as with more normal to adjacent tumor specimens, may be helpful to fully understand the role p38 plays in prostate cancer progression. Furthermore, this tissue microarray will serve as a useful resource for researchers to further study additional biomarkers in the future.

Annotating and filtering somatic variants in mesothelioma

Sophia Lee[1], Anish B. Chakka[2], Uma Chandran[2], Waqas Amin[2], Maureen Lyons-Weiler[3], Haroon Choudry[4], William LaFramboise[3], Michael J. Becich[2], David Bartlett[4]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, [3]Department of Pathology, University of Pittsburgh, [4]Division of Surgical Oncology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

E-mail: *Uma Chandran - chandran@pitt.edu
*Corresponding author

Context: Cancer genomes are characterized by somatic mutations, which may be involved in cancer initiation, progression, and pathogenesis. Next generation sequencing technologies have made it possible to identify these mutations in tumor samples. However, existing variant calling algorithms differ in their specificity and sensitivity.

Technology: The variant callers Samtools and Varscan2, the CLC Genomics Workbench, and Annovar were used in the analysis of whole exome sequence data from the SOLiD (Life Tech) platform.

Design: Using mesothelioma samples, we assessed the quality of the variant callers Samtools and Varscan2. Outputs from each caller were visually evaluated using a genome browser from CLC, and then the variants were annotated and filtered.

Results: Samtools produces a very large list of variants and appears to produce many false positives making the task of manual curation nearly impossible. Varscan2 developed specifically for finding cancer variants is better for cancer studies and produces a short list of variants to further annotate and curate. However, the overlap between the two variant callers is minimal. The somatic variants were annotated using annovar and filtered for exonic, nonsynonymous variants. This analysis was performed to identify novel somatic mutations and also mutations in known cancer genes including those that have been previously implicated in mesothelioma.

Conclusions: These findings suggest that the variant caller Samtools cannot be used to find somatic variants because it calls too many germline variants while Varscan2 is feasible to apply to cancer studies [Figure 1].

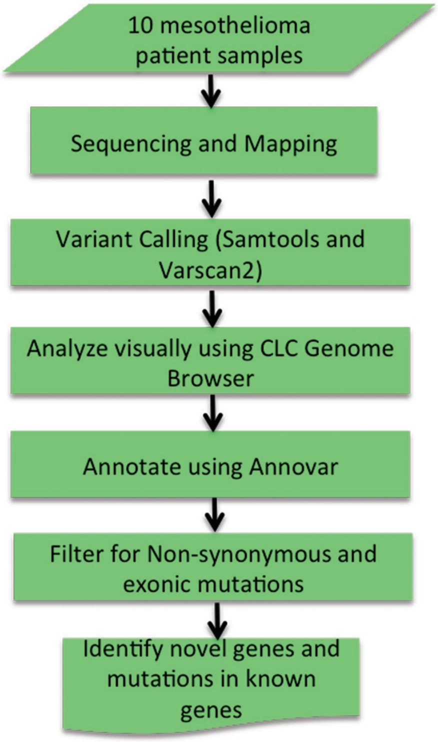

10 mesothelioma patient samples
Sequencing and Mapping
Variant Calling (Samtools and Varscan2)
Analyze visually using CLC Genome Browser
Annotate using Annovar
Filter for Non-synonymous and exonic mutations
Identify novel genes and mutations in known genes

Figure 1: Pipeline used to identify novel genes and mutations in known mesothelioma genes

Analysis of protein functions in cliques of protein interaction

Thomas Nash[1], Madhavi Ganapathiraju[2]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

E-mail: *Madhavi Ganapathiraju - madhavi@pitt.edu
*Corresponding author

Context: Protein-protein interactions provide clues about the functions of proteins. Analysis of the interactome,

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

i.e., the network of interactions, can be used to annotate functions of proteins by employing the principle of guilt-by-association. A clique in an interactome is a set of proteins where every protein in the clique interacts with every other protein. Here, we studied whether all the proteins in a clique have the same function.

Technology: We identified cliques using the Bron-Kerbosch algorithm whose implementation in MATLAB was available in open source. We used the DAVID online functional annotation table to analyze the functions of the proteins in the cliques.

Design: We downloaded the interactions from human protein reference database, processed the binary interactions to create a protein adjacency matrix. We analyzed this with the Bron-Kerbosch algorithm which outputs the cliques found in the network. Using DAVID functional analysis, we compared the functional similarity of proteins in cliques of different sizes to those in random gene sets of the same size. A box plot was used to represent the number of common functions of the proteins in the detected cliques.

Results: We found 9,006 cliques of size three, 3426 of size four, 1228 of size five, 476 of size six, 171 of size seven, 30 of size eight, and three cliques of size nine. We examined nine cliques each of sizes three to seven by manually studying them with DAVID. We determined that cliques have a higher functional similarity than randomly selected groups of proteins of the same size as there were a greater number of common gene ontology terms for the cliques than there were for the randomly selected groups of proteins. We found that the genes in cliques have more functional similarity than randomly selected groups of genes.

Conclusions: This approach may be used to develop an algorithm to predict the function of a gene where functions of other genes in its clique, if any, are known. This is more evident in cliques of a larger size.

Systems analysis focusing on adverse drug events within the nursing home setting

Brigitte Nguyen[1], Katrina Romagnoli[2], Richard D. Boyce[2]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

E-mail: *Richard D. Boyce - rdb20@pitt.edu
*Corresponding author

Context: Nursing homes are highly regulated environments that provide care to a large number of patients. Many healthcare issues in this setting can affect the safety of patients, such as adverse drug events. Understanding the nursing home healthcare system is vital to being able to address safety concerns and implement solutions. A systems analysis that combines data from qualitative interviews with retrospective electronic health records might help identify ways to reduce patient harm in the nursing home setting.

Design: We performed qualitative interviews with a variety of nursing home clinicians to explore their perceptions of medication safety in the nursing home. Interview questions were directed toward medication errors and adverse drug events. Situations discussed in the interviews were diagrammed into models based on interview transcripts. In addition, 3.5 years of medical records from 5 nursing homes (~5,000 patients) were queried to find relevant data that suggested harmful situations for patients based on problems mentioned in interviews.

Technology: The Dia program was used to create unified modeling language models of reported medication safety scenarios. SQL Queries were executed against the nursing home dataset that was stored in the Observational Medical Outcomes Partnership common data model. The dataset contained drug dispensing and minimum dataset data.

Results: Ten medication safety scenarios were identified from seven qualitative interviews. Reported errors included unintentional or inappropriate drug stops, as well as drug exposure beyond the required treatment period. Focusing on potential unintended drug stops attributable to care transitions, we found that, out of 788 patients who had a gap of 14 days or less from the nursing home, 33 people had an acute drug stop of either venlafaxine or paroxetine.

Conclusions: This research provides preliminary support for the feasibility of using dispensing and Minimum Dataset data to actively monitor for medication safety situations reported to occur in the nursing home setting.

Proposed system for two-way text messaging for discharged heart failure patients

Aditya Ravipati[1], Lingyun Shi[2], Mark Schmidhofer[3], Fuchiang Rich Tsui[2]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, [3]Presbyterian University Hospital of University of Pittsburgh Medical Center, Pittsburgh, PA, USA

E-mail: *Fuchiang Rich Tsui - tsui2@pitt.edu
*Corresponding author

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

Context: Postdischarge patient risk assessment is critical for reducing patient readmissions. Hospitals face a penalty for medicare patients readmitted within 30 days. We propose to develop a two-way text messaging system that can automatically assess risks of discharged heart failure patients from a hospital.

Technology: Microsoft Visio and Lucidchart were used to construct the flowchart for the whole system. Coding platforms for the PHP coding language, such as Notepad++, were used to code the rule engine of the system. My SQL was the open source database we used to store patient information and data. For patient, system communication SMS (text) messaging was used to ask questions and patient response. Finally, the system received a patient response through a Gmail account.

Design: The study comprises three parts: Creation of a questionnaire, development of flexible text messaging system, and creation of a survey form for user feedback of the proposed system. We consulted a cardiologist for what questions to ask for discharged patients. We developed a web interface that can send disease-specific questions to discharged patients in a sequence order and report adverse conditions to hospital staff who use the system. We also develop rules and simple natural language processing to process patient reported messages.

Results: We identified four key questions by consulting a cardiologist: (1) Are you currently taking all medicine as prescribed? (2) Are you gaining more than 3 pounds in last three days? (3) Do you have shortness of breath? and (4) Do you have swelling ankles? We developed rules for processing answers to the four questions and applied NegEx words for identifying negated words. We also developed a flowchart for the system which was used to code the rule engine for the questionnaire. We also developed the system and tested it with a smartphone using AT and T, which has proven to work the best with our system.

Conclusions: Given the popularity and simplicity of text messaging, we believe that our two-way text messaging system can be feasible to collect patient's reported conditions after the discharge. We expect the approach can be further expanded to other patient populations with other diseases.

Natural language processing tools for extracting opioid-related adverse drug events from clinical text

Kahmil Shajihan[1], Harry Hochheiser[2]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

E-mail: *Harry Hochheiser - harryh@pitt.edu
*Corresponding author

Context: Electronic Medical Records (EMRs) are a vast but largely unstructured source of information in a variety of medical fields. To use EMRs for research purposes, there is a need to structure this text. We applied natural language processing methods to identify adverse drug reactions to particular opioids.

Technology: We used an Unstructured Information Management Architecture (UIMA, uima.apache.org) workflow that searches through the de-identified EMRs using annotators looking for the pseudo-identification numbers, account numbers, dates of evaluation, drugs, and potential reactions.

Design: We used regular expressions to extract the identification numbers, account numbers, and dates. For the drug and reaction annotators, we created an external resource to determine what to extract. The external resource lists contained drug names and medical reactions that would focus the drug and reaction annotators on terms that could relate to an adverse drug event. To measure the sensitivity and precision of these annotators, we manually annotated documents and compared to the output of the annotators.

Results: The drug annotator had a 97.56% sensitivity but produced no false positives. The reaction annotator had 100% sensitivity but only 67.10% precision. The reaction annotator often picked up headings as hits.

Conclusions: The drug annotator functions well and avoids redundant annotations. The annotators withdrawing the dates, account numbers, and identification numbers also work very well due to the format of the EMRs. Drug and reaction processing could be improved with further refining of the external resources to maximize sensitivity. The precision of the reaction annotator was lowered by the annotators picking up headings as hits. To fix this problem, further investigation will be needed. We plan to combine the drug and reaction annotators to create an annotator that could detect potential adverse events described in a clinical note.

Consideration of potential clinical utility in MammaPrint validation studies

Marie Vater[1], Roger Day[2]

[1]University of Pittsburgh Cancer Institute Computer Science, Biology, and Biomedical Informatics (CoSBBI) Summer Academy, [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

E-mail: *Roger Day - Day01@pitt.edu
*Corresponding author

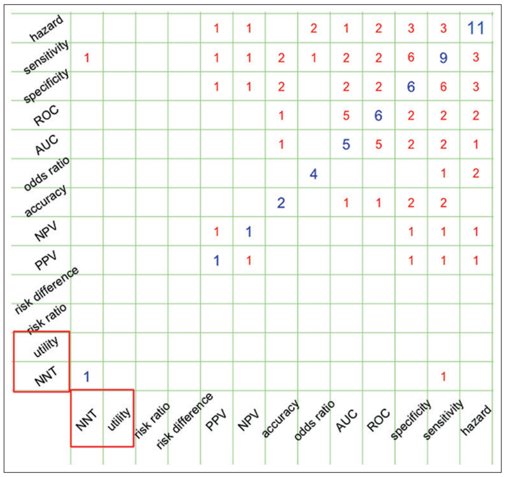
[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

Context: Despite efforts to exploit biomarkers for clinical application, the impact has been limited. A problem often cited is a deficiency in connecting studies to clinical utility. We aimed to assess this by determining the frequency, relevance, and placement of effect measures (EMs) and performance criteria relevant to patient decision-making as aspects of design, conducting, and/or results analyzing and reporting. We focused on scientific literature of MammaPrint (Agendia, Irvine, CA, USA), an expression-based biomarker developed for breast cancer patient prognostic determination. Validation studies (for overall effectiveness, additional value, and usage improvement) were surveyed. Inclusion criteria also regarded availability; only full, open-access articles in English qualified.

Design: A PubMed (Bethesda, Maryland, USA) search for mention of "MammaPrint" and/or "70-gene" in title and/or abstract was conducted (*n* = 155). Postexclusion and de-duplication, relevant and unique articles remained (*n* = 81). Further limitation left a subset for this preliminary study (*n* = 39).

Technology: Mendeley (Elsevier, London, England) was used to annotate use, type, and/or placement of: (1) power calculations for sample size and patient selection and classification, (2) EMs, (3) ethics-sensitive EMs reflecting ethical trade-offs required for clinical judgment, and (4) concrete deliberation of potential clinical utility in sections like discussion. Annotations were recorded using Microsoft Excel (Microsoft, Redmond, Washington, USA). Summaries were generated using RStudio and RMarkdown (RStudio, Boston, Maryland, USA).

Results: Only two articles presented a power calculation. 62% used at least one EM. The most common was hazard ratio (28%). Specificity was less frequent (15%) than sensitivity (23%). Predictive values, appropriate for prospective studies, were mentioned in only one article (3%). Figure 2 shows EM frequency (red boxes highlight ethics-sensitive EMs; numbers in blue are counts of individual EMs). Only 11 (28%) discussed the studies' clinical utility. Of those, 8 cited quantitative support in that discussion, but only 2 cited EMs. Only one (3%) considered potential clinical utility throughout the research process.



	NNT	utility	risk ratio	risk difference	PPV	NPV	accuracy	odds ratio	AUC	ROC	specificity	sensitivity	hazard
hazard					1	1		2	1	2	3	3	11
sensitivity	1				1	1	2	1	2	2	6	9	3
specificity					1	1	2		2	2	6	6	3
ROC							1		5	6	2	2	2
AUC							1		5	5	2	2	1
odds ratio								4				1	2
accuracy							2		1	1	2	2	
NPV					1	1					1	1	1
PPV					1	1					1	1	1
risk difference													
risk ratio													
utility													
NNT	1											1	

Figure 2: Frequency of effect measure co-occurrence

Conclusions: Lack of standardization of research performing and reporting was manifest in the haphazard use of different EMs and limited ethics-sensitive EMs. Those used were tangentially or, at best, minimally connected to specific clinically meaningful test performance. Tools and vocabulary to elicit relevant biomarker performance criteria are needed.

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

SUPPLEMENT LEGEND

Supplement 1: Final Computer Science, Biology, and Biomedical Informatics evaluation questionnaire

Interview #   Gender: Grade:

This interview will serve as part of an evaluation of the CoSBBI program. The responses that youprovide, which will be recorded, will be used to improve the program for the coming years in the form of a published editorial. Please respond openly and honestly, as the responses recorded today will be confidential and will not be traced back to you. The recording of your interview will be deleted upon the completion of the evaluation. We appreciate you taking the time to participate in this interview.

1. Why were you interested in applying to and attending CoSBBI and UPCI Summer Academy?

2. Rate your excitement level for CoSBBI and UPCI prior to starting. Why did you feel that way?

1	2	3	4	5
Not very excited	Not excited	Neutral	Excited	Very excited

3. Your expectations of the program were adequately met. Why or why not?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

4. The mentors were approachable and helpful. Why or why not?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

5. The instructors were approachable and helpful. Why or why not?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

6. What did you learn from your mentors that you feel will help you down the road, not just in biomedical informatics?

7. What did you learn from your instructors that you feel will help you down the road, not just in biomedical informatics?

8. You enjoyed the lectures and found them to be engaging. Why?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

9. Which lectures stood out as particularly enjoyable? Why?

10. Which lectures stood out as particularly unenjoyable? Why?

11. You found the research hour to be interesting. Why or why not?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

12. Which topic would you have liked to learn about more in depth from the lectures? Why?

13. You found your research project to be interesting and enjoyable. Why or why not?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

14. What would you improve or do differently about your research project if you could do it over?

[Downloaded free from http://www.jpathinformatics.org on Wednesday, January 31, 2018, IP: 130.49.179.43]

15. What are your perceptions of minorities in the sciences? Such as women and underrepresented racial minorities.
16. How do you think we can better attract these groups into the field?
17. How would you rank this academy in terms of a summer career preparation experience?
    Among the best? Worst? Why? Which do you consider the best?
18. Did you feel that the academy helped you in your process of choosing a career path? Why or why not?
19. Has CoSBBI influenced your interest in the field of biomedical informatics? Why or why not?
20. What is the most important thing that you will take away from this summer experience?
21. You found the social and educational events enjoyable. Why or why not?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

22. Which was your favorite? Least favorite?
23. You found the program to be a useful experience that you would recommend to others. Why or why not?

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly Agree

24. How could the academy be improved?