A Pan-Cancer Authorship Network Analysis

Tiffany Wu
Department of
Biological Sciences
Mount Holyoke College
South Hadley, MA

Elizabeth Sigworth, BA
Department of
Biostatistics
Vanderbilt University
Nashville, TN

Xuanyi Li, BS
School of Medicine
Vanderbilt University
Nashville, TN

Samuel M. Rubinstein, MD
Department of
Hematology/Oncology,
School of Medicine
Vanderbilt University
Nashville, TN

Jeremy L. Warner, MD, MS
Department of
Hematology/Oncology &
Biomedical Informatics
Vanderbilt University
Nashville, TN

I. Introduction

Progress in the field of hematology/oncology is mediated primarily through the publication of practice-changing clinical trials. Individual researchers typically specialize in their research and publish in one subfield. We established measures of an author's impact as well as the diversity of their publications across cancer subfields to generate a weighted network, which provides a novel method for visualizing an author's significance and offers insight into the links between researchers across this rapidly evolving field.

II. Methods

Our dataset draws from the prospective clinical trial literature cited on HemOnc.org, a wiki-based website primarily intended for hematology/oncology professionals. It spans roughly 20,000 authors across over 120 cancer subtypes.

A. Gini Index

The Gini Index classically measures income disparity in a country by evaluating the inequality of values among income level. We repurposed it to measure an author’s subfield diversity using the distribution of their publications across 12 cancer subfields (e.g., thoracic oncology; breast cancer; lymphoma). A higher Gini coefficient indicates more inequality, or specialization.

B. Global Impact Score

Scores were assigned based on author position and the impact factor of the journal of publication, with first or last authors and higher tier journals having more weight.

III. Graphical Results

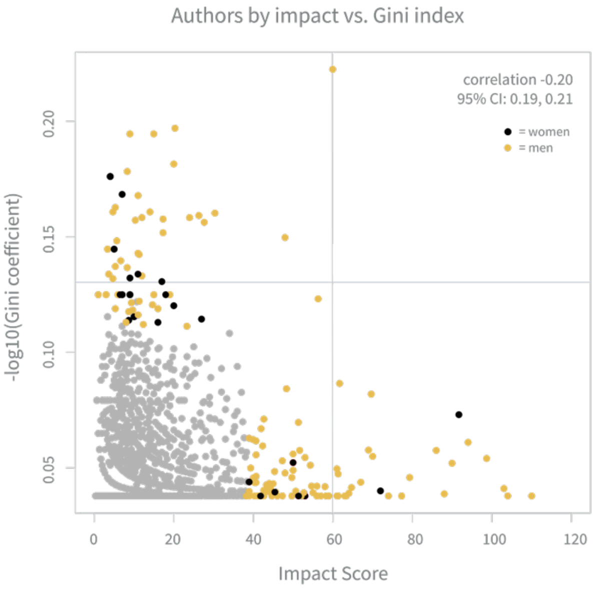

Figure 1. Plot of authors by impact and Gini coefficient. Top 100 by impact and by Gini are colored by gender. Note that the y-axis is flipped (higher placement = lower Gini score).

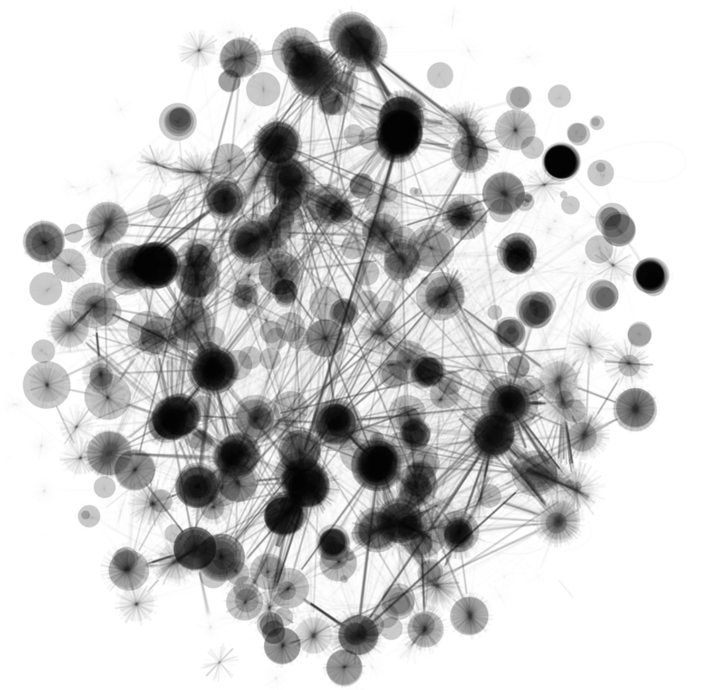

Figure 2. Graph of network with authors who publish in more subfields emphasized with increased opacity. Larger node size indicates a greater impact score.

IV. Discussion

More collaboration across subfields has a statistically significant correlation with a higher median impact (correlation=-0.20 *for all authors*, p-value < 0.0001; 95% confidence interval of -0.19, -0.21), as shown in ***Figure 1***. Additionally, the visual gender distribution of the *top 100 authors* by each measure suggests a significant disparity (colored points). Of the top hundred authors by impact, 91% are men, while they compose 68% of the 100 with lowest Gini index. The odds that a woman is in the top 100 authors by Gini coefficient (i.e., the lowest Gini indices) are 4.64 times the odds that she is in the top 100 by impact (p-value < 0.0001). Strikingly, authors with low Gini coefficient are responsible for the overall network structure shown in ***Figure 2***; without them, the network would fall apart, leaving only a collection of siloed subgraphs.

Future work will focus on the ratio of in-links, or connections to authors who are classified as being in the subfield, to out-links, or connections to authors who are classified as being in another subfield, while also considering the average Gini coefficient of the author’s collaborators. We further expect to look into the correlation of Gini coefficient with median publication to see if this aligns with the traditional narrative of cancer history, e.g., as chronicled in Siddhartha Mukherjee’s *The Emperor of All Maladies*, as well as evaluate temporal trends in network structure and examine additional attributes of authorship, e.g., years in the field, institutional affiliation, and so forth.