Research Synergy and Drug Development: Bright Stars in Neighboring Constellations

Samet Keserci[1], Eric Livingston[2], Lingtian Wan[1,¤], Alexander R. Pico[3], & George Chacko[1*]

1 Netelabs, NET ESolutions Corporation, McLean, VA, USA
2 Elsevier Research Intelligence, Elsevier Inc., Bethesda, MD, USA
3 Gladstone Institutes, University of California San Francisco, San Francisco, CA, USA

¤Current Address: Facebook Corporation, Menlo Park, CA, USA
*Correspondence: george@nete.com

Abstract

Drug discovery and the subsequent availability of new breakthrough therapeutics or ‘cures’ are compelling examples of societal benefit from advances in research. Such advances are invariably collaborative, involving the contributions of many scientists building upon prior theory and experiment. Data mining of public and commercial data sources coupled with analysis of ‘cure networks’ is a scalable digital methodology for assembling and analyzing the scientific history of these advances. The methodology for assembling is extensible beyond cure networks and such studies support (i) efficiency in exploring science history (ii) an improved documenting and understanding of collaboration (iii) portfolio analysis, planning and optimization (iv) communication of the societal value of research. In addition, they help make the scientific environment more citizen-accessible. Building upon techniques for single cure networks, we have conducted a case study of five anti-cancer therapeutics for the purpose of exploring the relationship across these networks. We have enriched the content of these networks by annotating them with information on research awards as well as peer review that preceded these awards. Applying retrospective citation discovery, we have identified a core set of publications cited in the networks for all five therapeutics and additional intersections in pairwise interactions between networks.

Introduction

Data mining of public data sources coupled with network analysis allows the assembly and quantitative description of research discoveries that were influential in the development of a breakthrough therapeutic or ‘cure’. The set of scientific publications, clinical trials, patents, and regulatory approvals, linked to each other by citation or assignment, that documents progress from basic research to a cure is termed a ‘cure network’ [1]. Key assumptions in constructing these networks are that the references found in relevant documents are appropriate citations of new knowledge relevant to a given cure and that a further retrospective round of citation discovery will uncover previous influential discovery (*ibid*). Williams and colleagues have elegantly demonstrated the feasibility of this approach using, as case studies, ivacaftor and ipilimumab, approved for the treatment of cystic fibrosis and melanoma respectively (*vide supra*). These authors observed that ‘the nature of a cure discovery network is complex and fundamentally collaborative', noting in the case of ivacaftor, that at least 7,067 scientists with 5,666 unique affiliations contributed to ivacaftor-related research over a period greater than 100 years.

Such studies provide evidence for the broad collaborative platform of basic and translational research underlying major scientific advances such as cures for diseases, support strategic communications to oversight bodies, and communication of the societal value of research [2,3]. Extending this digital methodology to study additional cures as well as significant research advances in general is a logical next step. Ascertaining the nature of the interactions, if any, between networks, is also of considerable interest since it would support an understanding of collaboration across networks as well as common features of science networks.

To address these questions, we have conducted case studies of a cluster of five FDA-approved therapeutics for cancer. We have extended the original approach to (i) include a data from a commercially available bibliographic database with disambiguated author identifiers (ii) modified the original data mining methods and network metrics (iii) added new node types to include grants and peer review data (iv) refactored the network analysis code of Williams et al. for better performance. (v) mapped publications and authors across all five networks.

We present the results of these case studies as a step towards mapping the universe of FDA-approved drugs and biologicals. Beyond drug discovery and cures, however, this simple digital methodology can be used to support efficient explorations of science history, gain an understanding of collaboration across domains, enrich portfolio analysis, planning and optimization, enable communications of the societal value of research, and help make the scientific environment more citizen-accessible.

Materials and Methods

A set of five anti-cancer therapeutics, three drugs and two biologicals, approved for use in humans by the Food and Drug Administration was selected for this study (Table 1). Imatinib and Sunitinib are tyrosine kinase inhibitors, Nelarabine is a nucleoside analog, and Ramucirumab and Alemtuzumab are humanized antibodies that target cell surface receptors. For each of these therapeutics, a set of relevant scientific publications was constructed as in Williams et al. [1] but with specific modifications detailed below. An allowance of 2 months was made for 'publication lag' when assembling referenced material. For example, if a therapeutic was approved on Jan 1, 2017, documents published on or before March 31, 2017 were included. For each of the five therapeutics, a first-generation list of PubMed IDs (citing_pmid) was harvested from the five different data sources (below).

Therapeutic	FDA Approval Date	Unique Identifier	US Patent	Publication Date
Alemtuzumab	May 2001	BLA: 103948	US5846534	Dec 1998
Imatinib	May 2001	NDA: 021335	US5521184	May 1996
Nelarabine	Oct 2005	NDA: 021877	US5424295	Jun 1995
Ramucirumab	Apr 2014	BLA: 125477	US7498414	Mar 2009
Sunitinib	Jan 2006	NDA: 021938	US6573293	Jun 2003

Table 1. Case Studies of Five Anti-Cancer Agents Five anti-cancer therapeutics, with FDA approval dates ranging from 2001 to 2014, were selected as case studies. The active ingredient for each of these five therapeutics is listed in column 1. While multiple patents are typically associated with a drug or biological, the single US patent number displayed represents the primary invention that preceded approval of the therapeutic. The publication date for each patent is listed in the last column.

Clinical trials The national clinical trials database (clinicaltrials.gov) was searched for clinical trials of the five therapeutics that completed on or the data of FDA approval. Both cited references and publications from these clinical trials were collected if they were published within the approval date plus two months period. PubMed was also searched with the unique identifier (NCT number) of any clinical trials that were identified to capture publications associated with the clinical trials that were not displayed in clinical trials.gov. To capture clinical trials publications that are not available in clinicaltrials.gov, PubMed was searched using the therapeutic name as keyword, publication type as "clinical trial", and an appropriate date restriction as in searches of clinicaltrials.gov. For example, the search term ((("alemtuzumab"[Supplementary Concept] OR "alemtuzumab"[All Fields]) OR ("alemtuzumab"[Supplementary Concept] OR "alemtuzumab"[All Fields] OR "campath"[All Fields])) AND ("1900/0101"[PDAT] : "2001/07/31"[PDAT])) AND "clinical trial"[Publication Type] was used to identify publications of clinical trials for Alemtuzumab.

FDA documents The drugs@fda website [6] was searched for each of the five therapeutics and cited references in the medical review document were manually extracted and matched to pmids. FDA Approval Summaries, articles published in journals by FDA staff, were available for all five therapeutics and contain cited references. If the published date of a citation exceeded the approval date plus two months, the publication was not included.

Patents For each therapeutic a single patent was identified that best represented the most relevant invention to the therapeutic at hand. Identification of this patent was performed using multiple web sources. The US patent number was then used to identify the patent and the non-patent citation list from Google Patents [8] was manually processed by searching PubMed for appropriate pmids. While we developed a citation matching tool for non-patent literature for this purpose, the precision and recall rates from manual searches was far higher and were used to generate the data in this study.

Post-approval literature reviews Review articles published after a therapeutic's approval by the FDA are independent assessments of the the development of a therapeutic. Accordingly, PubMed was searched for review articles of a therapeutic that were published between the date of FDA approval and one year following the date of approval and cited references in these reviews were extracted using PubMed and Scopus.

Pre-approval literature searches Literature searches were performed using PubMed with a date range of 1900/01/01 to two months post-FDA approval. For example, the search term ((alemtuzumab) OR campath) AND ("1900/01/01"[Date - Publication] : "2001/07/31"[Date - Publication]) was used to retrieve articles of interest relevant to alemtuzumab.

Citing_pmids from the five different sources above were combined and deduplicated. Using the Scopus database and its APIs, these citing_pmids were mapped to ScopusIds (citing_sid) and a second generation of cited references (cited_sid) was extracted., These cited_sids, were in turn, mapped back to pmids (cited_pmid). Whereas mapping between PubMed and Scopus identifiers at the citing_pmid and citing_eid stage resulted in 1% or less information loss, mapping at the cited_eid to cited_pmid resulted in 15-20 % information loss. Accordingly the Scopus data was used as the backbone of the publication component of the network and cited_pmid was treated as an annotation layer. These observations are summarized in Table 2 and include the count of null values when mapping from citing_sid to cited_sid.

Citing and cited pmids were both mapped to NIH grants and peer review panels using information made publicly available by NIH through NIH ExPORTER. Since Special Emphasis Panels vary in scientific interests from year to year, we restricted our mapping to chartered study sections, which have defined scientific interests and

Therapeutic	citing_pmid count	citing_eid count	cited_eid count	cited_pmid count
Alemtuzumab	599	587 (1%)	8840 (2%)	7071(20%)
Imatinib	1380	1373(1%)	27326(1%)	23340(17%)
Nelarabine	104	104(0%)	2476(1%)	1990(20%)
Ramucirumab	1820	1804(1%)	48587(0%)	40973(19%)
Sunitinib	1512	1509(0%)	33895(0%)	28661(15%)

Table 2. Citation Counts and Mapping Between Bibliographic Databases Five anti-cancer therapeutics were selected as case studies. A foundational set of references (citing_pmid) was assembled for each therapeutic from patents, clinical trials, regulatory documents, and the scientific literature. Citing_pmids were mapped to Scopus identifiers (citing_eid), which were used, in turn, to retrieve literature cited by this foundational set (cited_eid). Cited_eids were mapped back to PubMed identifiers (cited_pmid).The number of identifiers at each stage of the mapping process is shown along with percentage loss (in parentheses) when mapping across PubMed and Scopus or due to null values in the cited_sid field

relatively stable reviewer membership. The resultant data were modeled as hierarchical networks and analyzed using metrics based on network topology. We calculated the propagated in-degree rank (PIR) metric of Williams [1]. PIR represents the sum of weighted citation scores for all articles in a network that can be attributed to an author. In addition to computing PIR for all authors in each network, we also combined the citation data for all five networks and computed a globalPIR score, which was normalized to the sum of PIR scores within each network.

We also calculated the RBR metric of Williams et al. (*ibid*) with some modifications. The RBR metric is intended to represent the fraction of a researcher's output that is in a network and is defined as the ratio of the number of publications in network to the number of publications in a background dataset for an author. In its original specification, the background dataset for RBR was constructed by keyword searches of PubMed. A potential weakness of this keyword based approach is that keywords do not effectively capture the field or the total output of an author even if multiple background samples are taken. Therefore, we created two new variants of the RBR; network RBR (nRBR) and document-based RBR (dRBR). nRBR uses all publications in our set of five therapeutics as background and dRBR takes advantage of the Scopus author_id to capture the total output of an author and use it as background. Thus, nRBR and dRBR normalize a researcher's in-network contributions to backgrounds based on total network and total researcher productivity respectively.

Etiam eget sapien nibh.

Nulla mi mi, Fig 1 venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, S1 Video vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Fig 1. Bold the figure title. Figure caption text here, please use this space for the figure panel descriptions instead of using subfigure commands. A: Lorem ipsum dolor sit amet. B: Consectetur adipiscing elit.

Results and Discussion

Network Features

Intersection Across Networks

this is a

Supporting awards and peer-review

Future Directions

alem(A)	imat(I)	nela(N)	ramu(R)	suni(S)	combination	intersection_count	no_of_drugs
X	X	X	X	X	AINRS	**14**	5
X	X		X	X	AIRS	**107**	4
X	X	X	X		AINR	28	4
X	X	X		X	AINS	25	4
	X	X	X	X	INRS	22	4
X		X	X	X	ANRS	16	4
	X		X	X	IRS	**1762**	3
X	X		X		AIR	231	3
X	X			X	AIS	211	3
X			X	X	ARS	156	3
X	X	X			AIN	81	3
X		X	X		ANR	56	3
	X	X		X	INS	49	3
	X	X	X		INR	44	3
		X	X	X	NRS	40	3
X		X		X	ANS	32	3
			X	X	RS	**7442**	2
	X			X	IS	5415	2
	X		X		IR	2507	2
X	X				AI	1240	2
X			X		AR	448	2
X				X	AS	359	2
X		X			AN	334	2
	X	X			IN	232	2
		X	X		NR	115	2
		X		X	NS	111	2
			X		R	**49006**	1
				X	S	34257	1
	X				I	27706	1
X					A	8978	1
		X			N	2498	1

Table 3. Intersecting Publications Across Networks Intersections were calculated for the set of publications associated with each network. Both citing and cited references were included in each set and Scopus identifiers were used to to minimize information loss (Table 2). Intersection counts are shown for all possible combinations of the five therapeutics. The largest intersection count in each combination group is shown in boldface. Therapeutic names are abbreviated as follows: Alemtuzumab as alem(A), Imatinib as imat(I), Nelarabine as nela(N), Ramucirumab as ramu(R), and Sunitinib as suni(S).

No idea what to say...

LOREM and IPSUM Nunc blandit a tortor.

3rd Level Heading.

Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

1. react
2. diffuse free particles
3. increment time by dt and go to 1

Sed ac quam id nisi malesuada congue.

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

- First bulleted item.
- Second bulleted item.
- Third bulleted item.

Discussion

Nulla mi mi, venenatis sed ipsum varius, Table 1 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero [3].

Conclusion

CO_2 Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Appendix.

Supporting Information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

$$\mathrm{P_Y} = \underbrace{H(Y_n) - H(Y_n|\mathbf{V}_n^Y)}_{S_Y} + \underbrace{H(Y_n|\mathbf{V}_n^Y) - H(Y_n|\mathbf{V}_n^{X,Y})}_{T_{X\to Y}}, \tag{1}$$

References

1. Williams RS, Lotia S., Holloway AK, Pico AR. From Scientific Discovery to Cures: Bright Stars within a Galaxy. Cell. 2015 Sep; 163:21–23

2. Lauer MS. PCSK9 Inhibitors: Lots of Work Done, Lots More to Do. Ann Intern Med. 2016 Mar; 164(9):624-625.

3. Wang M., Zhao Y.., Zhang B. Efficient Test and Visualization of Multi-Set Intersections. Sci. Rep. 2015 Nov; doi:10.1038/srep16923

4. O'Shea JJ, Kanno, Y., Chan AC. In Search of Magic Bullets: The Golden Age of Immunotherapeutics Cell. 2014 Mar; 157:227-240

5. Maldame, J. The Importance Of The History Of Science In Intellectual Formation Scripta Varia 2002 104:237-248

6. Federal Drug Administration. Drugs@FDA: FDA Approved Drug Products https://www.accessdata.fda.gov/scripts/cder/daf/

7. National Institutes of Health. NIH ExPORTER https://exporter.nih.gov/

8. Google Corporation Google Patents https://patents.google.com/

9. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive increases in the resistance of Drosophila to viral infection through a transposon insertion followed by a Duplication. PLoS Genet. 2011 Oct;7(10):e1002337.