Heliyon

Received:
16 September 2017
Revised:
15 October 2017
Accepted:
26 October 2017

Dataset link:
Data associated with this study has been deposited at Mendeley Data under the accession number DOI: 10.17632/ysh53v7gpz.4

Cite as: Samet Keserci, Eric Livingston, Lingtian Wan, Alexander R. Pico, George Chacko. Research synergy and drug development: Bright stars in neighboring constellations. Heliyon 3 (2017) e00442. doi: 10.1016/j.heliyon.2017.e00442

Research synergy and drug development: Bright stars in neighboring constellations

Samet Keserci [a], **Eric Livingston** [b], **Lingtian Wan** [a,1], **Alexander R. Pico** [c], **George Chacko** [a,*]

[a] *NETE Labs, NET ESolutions Corporation, McLean, VA, USA*

[b] *Research Intelligence, Elsevier Inc., New York, NY, USA*

[c] *Gladstone Institutes, San Francisco, CA, USA*

* Corresponding author.
E-mail address: netelabs@nete.com (G. Chacko).

[1] Current address: Facebook Inc., Menlo Park, CA, USA.

Abstract

Drug discovery and subsequent availability of a new breakthrough therapeutic or ‘cure’ is a compelling example of societal benefit from research advances. These advances are invariably collaborative, involving the contributions of many scientists to a discovery network in which theory and experiment are built upon. To document and understand such scientific advances, data mining of public and commercial data sources coupled with network analysis can be used as a digital methodology to assemble and analyze component events in the history of a therapeutic. This methodology is extensible beyond the history of therapeutics and its use more generally supports (i) efficiency in exploring the scientific history of a research advance (ii) documenting and understanding collaboration (iii) portfolio analysis, planning and optimization (iv) communication of the societal value of research. Building upon prior art, we have conducted a case study of five anti-cancer therapeutics to identify the collaborations that resulted in the successful development of these therapeutics both within and across their respective networks. We have linked the work of over 235,000 authors in roughly 106,000 scientific publications that capture the research crucial for the development of these five therapeutics. Applying retrospective citation discovery, we have identified a core set of publications cited in the networks of all five therapeutics and additional intersections in combinations of networks. We

https://doi.org/10.1016/j.heliyon.2017.e00442

have enriched the content of these networks by annotating them with information on research awards from the US National Institutes of Health (NIH). Lastly, we have mapped these awards to their cognate peer review panels, identifying another layer of collaborative scientific activity that influenced the research represented in these networks.

Keywords: Information science, Cancer research

1. Introduction

Data mining of public data sources coupled with network analysis enables the 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 the progress of concepts from basic research to a cure has been termed a ‘cure network’ [1]. Science history studies in general and network approaches have been proposed to enable studies of knowledge diffusion across disciplines, scientific interests, culture, and time [2, 3]. Such studies also (i) provide evidence for the broad collaborative platform of basic and translational research underlying major scientific advances such as cures for diseases [4] support strategic communications that help communicate the societal value of research. The understanding of a therapeutic network, when coupled with information from clinical use of a therapeutic, also enables a recursive learning of the pathogenesis of the disease it is being used to treat, as has been noted for the burgeoning field of immunotherapeutics [5].

Williams and colleagues have elegantly demonstrated the feasibility and value of data mining and network analysis using, as case studies, ivacaftor and ipilimumab, approved for the treatment of cystic fibrosis and melanoma respectively (*vide supra*). Key assumptions in constructing these networks were that 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 reveal previous influential work. They 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-relevant research over a period greater than 100 years. These authors also suggested that thoughtful metrics derived from this concept could inform decision making by funders.

In this study, we document the collaboration networks underlying five FDA-approved therapeutics for cancer. Building upon prior art for single networks we (i) extend single network analysis to map publications and authors across multiple networks

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 unique identifier for each therapeutic is an FDA assigned NDA or BLA number. 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.

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

(ii) include information on research awards and peer review of grants (iii) include enriched data from a commercially available bibliographic database with disambiguated author identifiers, and (iv) incorporate modified network metrics and data mining methods. We observe collaboration that extends across networks and describe the role of funding and peer review in sustaining a system of layered collaborative activity in scientific discovery. By studying additional cures and research advances, we also proceed towards scaling from single case studies to mapping the entire domain of drug development with the expectation that such knowledge will be beneficial in planning, resource allocation and optimization of drug development research. We present the results of this case study to demonstrate a framework that can be easily modified by other researchers to generate new datasets or to complement existing ones.

2. Materials and methods

Five anti-cancer therapeutics, three drugs and two biologicals, approved for use in humans by the Food and Drug Administration were selected for this study (Table 1). Imatinib [6] and Sunitinib [7] are tyrosine kinase inhibitors, Nelarabine [8] is a nucleoside analog, and Ramucirumab [9] and Alemtuzumab (Campath) [10] are humanized antibodies that target the CD52 and vEGFR-2 cell surface receptors respectively. 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.

Clinical trials. The national clinical trials database (clinicaltrials.gov) was searched for clinical trials of the five therapeutics that completed by the data of FDA approval by searching for the therapeutic name in the intervention field. Both cited references and publications from these clinical trials were collected if they were published within the approval date plus two months. To capture publications associated with the clinical trials that were not displayed in clinicaltrials.gov, PubMed was also searched with the unique identifier (NCT number) of any clinical trials that were identified. To capture publications of clinical trials not registered in clinicaltrials.gov, PubMed was searched using the therapeutic name (imatinib, alemtuzumab, sunitinib, ramucirumab, and nelarabine) as keywords without using synonyms, 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 [11] was searched for each of the five therapeutics. The medical review document located under Approval Date(s) and History, Letters, Labels, Reviews was copied. Cited references in the medical review document were manually extracted and matched to pmids using, as search terms in the PubMed GUI, text strings from the citation that, typically, consisted of the last name of the first author, 3–5 words of the title, the year of publication, and journal name. FDA Approval Summaries published in journals by FDA staff, were available for all five therapeutics and contain cited references. If the published date of a cited reference in an Approval Summary exceeded the approval date plus two months, the publication was not included.

Patents. Using a combination of web searches, Google Patents, and the scientific literature, a single patent was subjectively identified that best represented the most relevant invention for the therapeutic being considered. Identification of this patent was performed using multiple web sources. The US patent number was then used to identify the patent. For example, US5521184, imatinib, was assigned to the Ciba-Geigy Corporation in 1996 listing Zimmermann as inventor. The non-patent citation list for a patent was copied from Google Patents [12] and manually processed by searching the PubMed GUI for appropriate pmids using text strings from the citation that, typically, consisted of the last name of the first author, 3–5 words of the title, the year of publication, and journal name. Returned hits were inspected for matches to the original citation in Google Patents and accepted only in the event of a high-degree of confidence in correctness.

Post-approval literature reviews. Review articles published after a therapeutic's approval by the FDA are independent studies of the development of a therapeutic. Accordingly, PubMed was searched for review articles on these five therapeutics that were published between the date of FDA approval and a year following the date of approval. Cited references in these reviews were extracted using PubMed and Scopus. The review articles themselves were not included.

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. For each of the five therapeutics, a first-generation list of PubMed identifiers (citing_pmid) was harvested from the five different data sources.

PubMed and Scopus. Citing_pmids from the five different sources above were combined and deduplicated. Using the Scopus database and its APIs, the manually-generated list of pmids taken from the five sources mentioned (Clinical Trials, FDA Documents, Patents, Post-Approval Literature, and Pre-Approval Literature Searches) were searched in Scopus, using the basic Scopus Search API, to arrive at a list of Scopus IDs (citing_sid) for the publications. The Scopus Abstract Retrieval API was then used to retrieve a more comprehensive record for each of the SIDs comprising that list of publications. Next, for each of these publication records (citing_sid), we used the Scopus Author Retrieval API to retrieve a full record for each unique author in the publication set. We also used the Abstract Retrieval API to collect records for each of the publications cited by the first generation of publications. This set of cited publications is the cited_sid set. Using the same Author Retrieval API, we then gathered data for each of the unique authors affiliated with the cited_sid publications. Completion of the process yields two sets of publications, citing_sid and cited_sid, with citation links between them and full information on all authors for both generations. Finally, for each author in the study, we used the standard Scopus Search API once more to retrieve a smaller record for every publication affiliated with them in Scopus, in order to tally their overall publication output. While author records in Scopus have overall publication counts as part of the record, by manually downloading each of them, we can store and count them by type (i.e. article, book chapter, Editorial, review, etc.). This allowed us to more precisely arrive at publication totals for only those publication types that are relevant for this study. For the gRBR metric (below), we counted only article and article in press when computing total researcher productivity.

Synonyms were partially accounted for in the PubMed Advanced Search feature. A search for Gleevec was automatically translated to "imatinib mesylate"[MeSH Terms] OR ("imatinib"[All Fields] AND "mesylate"[All Fields]) OR "imatinib mesylate"[All Fields] OR "gleevec"[All Fields]. In other cases, synonyms were added into the search by the authors. For example, Campath for Alemtuzumab. ((alemtuzumab) OR campath) AND ("1900/01/01"[Date - Publication] : "2001/07/31"[Date - Publication]) Whereas mapping between PubMed and Scopus identifiers at the citing_pmid and citing_sid stage resulted in 1% or less information loss, mapping at the cited_sid to cited_pmid resulted in a loss of roughly 15–20%

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 (Materials and Methods). Citing_pmids were mapped to Scopus identifiers (citing_sid), which were used, in turn, to retrieve cited publications (cited_sid). Cited_sids 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.

Therapeutic	citing_pmid count	citing_sid count	cited_sid 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%)

of target records. Accordingly the Scopus data was used as the backbone of the publication component of the network and the cited_pmid information was treated as an annotation layer. These observations are summarized in Table 2.

Both citing and cited pmids were mapped to NIH grants and peer review panels (study sections) using public information available through NIH ExPORTER [13]. Thus, we enriched our network data by identifying those study sections associated with the awards that supported publications in our networks.

Networks and network calculations. The resultant data were modeled as networks and analyzed using metrics based on network topology. We calculated the propagated in-degree rank (PIR) and ratio of basic rankings (RBR) metrics of Williams [1]. PIR represents the sum of aggregated citation scores (first and second degree only) 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 network PIR (nPIR) score, which was normalized to the sum of individual PIR scores within each network as the PIR PartitionRatio (PPR) as a way to measure inter-network influence. RBR 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 it does 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 global-based RBR (gRBR). nRBR uses all publications in our set of five therapeutics as background and gRBR takes advantage of the Scopus author_id to capture the total article output of an author as background. Thus, nRBR and gRBR normalize a researcher's in-network contributions to backgrounds based on total network and total researcher productivity respectively.

https://doi.org/10.1016/j.heliyon.2017.e00442

Propagated Indegree Rank (PIR). In this study, we examine five therapeutics $d_1, d_2, \ldots, d_5$, and their networks $N_1, N_2, \ldots, N_5$. The nodes of network N_i are the publications associated to the therapeutic d_i, and the directed edges of N_i are obtained by the global network $\mathcal{G}$. Thus, we include a directed edge between publications x and y if and only if x cites y in $\mathcal{G}$. Hence, N_i is a simple graph (no parallel edges and no self-loops).

We define network $\mathcal{N}$ to be the graph-theoretic union of the networks $N_1, N_2, \ldots, N_5$ (i.e., $\mathcal{N} = \cup_i N_i$). Thus, the nodes of the network $\mathcal{N}$ are the nodes that appear in at least one network N_i, and we include a directed edge between publications x and y if and only if x cites y in at least one of the networks N_i; hence, $\mathcal{N}$ is a simple graph (no parallel edges and no self-loops).

Let $\mathfrak{n}$ denote some selected network $\mathcal{N}_i$, let $c_\mathfrak{n}(\mathfrak{p})$ be the citation score of publication $\mathfrak{p}$ in $\mathfrak{n}$, and let $C^{\mathfrak{n}}_{\mathfrak{p}}$ be the set of publications in $\mathfrak{n}$ that cite $\mathfrak{p}$.

We define the aggregated citation count for $\mathfrak{p}$ within network $\mathfrak{n}$, denoted by $ac_\mathfrak{n}(\mathfrak{p})$, by

$$ac_\mathfrak{n}(\mathfrak{p}) = c_\mathfrak{n}(\mathfrak{p}) + \sum_{g \in C^{\mathfrak{n}}_{\mathfrak{p}}} c_\mathfrak{n}(g).$$

Let $\mathcal{A}^{\mathfrak{n}}_a$ be the set of publications for an author $\mathfrak{a}$ in $\mathfrak{n}$. Then the PIR score for $\mathfrak{a}$ in network $\mathfrak{n}$ is defined by

$$pir_\mathfrak{n}(\mathfrak{a}) = \sum_{p \in \mathcal{A}^{\mathfrak{n}}_a} ac_\mathfrak{n}(\mathfrak{p})$$

Hence,

$$pir_\mathfrak{n}(\mathfrak{a}) = \sum_{p \in \mathcal{A}^{\mathfrak{n}}_a} \left[c_\mathfrak{n}(p) + \sum_{g \in C^{\mathfrak{n}}_{\mathfrak{p}}} c_\mathfrak{n}(g) \right]$$

Next we define the nPIR score of author $\mathfrak{a}$ within network $\mathcal{N}$ (denoted by $nPIR(\mathfrak{a})$) to be $pir_{\mathcal{N}}(\mathfrak{a})$; in other words it is the *pir* score based on the network $\mathcal{N}$.

We define the PIR partition ratio (PPR) of author $\mathfrak{a}$ (denoted by $ppr(\mathfrak{a})$) to be

$$ppr(\mathfrak{a}) = \frac{nPIR(\mathfrak{a})}{\sum_{i=1}^{5} pir_{N_i}(\mathfrak{a})}$$

There are cases where PPR can be greater than 1.

Ratio of Basic Rankings (RBR).

- nRBR (i.e., network RBR) is the ratio of an author's publication count in a given network $\mathfrak{n}$ to the total publication count for that author in $\mathcal{N}$. Thus, nRBR depends on both the author $\mathfrak{a}$ and the given network $\mathfrak{n}$, and is denoted by $nRBR(\mathfrak{a}, \mathfrak{n})$.

- gRBR, or global RBR, is the ratio of an author's count of publications in network $\mathfrak{n}$ to the author's total publication count in the global network $\mathcal{G}$, and is denoted by $\text{gRBR}(\mathfrak{a}, \mathfrak{n})$.

$\text{nRBR}(\mathfrak{a}, n)$ and $\text{gRBR}(\mathfrak{a}, n)$ are both ratios with the same numerator but with different denominators, and $0 \leq \text{nRBR}(\mathfrak{a}, \mathfrak{n}) \leq \text{gRBR}(\mathfrak{a}, \mathfrak{n}) \leq 1$.

Analysis. All data used in this study were acquired exclusively from the sources listed above. Data used to generate the figures and tables in this study are available in a Mendeley Data repository [14]. Computations were performed on infrastructure owned or leased by NETE Solutions, Elsevier, or the Gladstone Institutes. Code and scripts used in this study were written in Java, Python, and R and are archived on a publicly accessible Github repository [15]. The publicly available codes of the Williams study [16] were used as the basis for designing the codes used in our study. The previous codes were designed to generate graph objects and make use of graph methods such as breadth-first search (BFS) and depth-first search (DFS for graph traversal). In our approach, we used basic data structures like hash maps, hash sets, lists and aggregations and enriched the first generation set of references with data from Scopus, which indexes more scientific journals [17], cited references from post-approval literature reviews, cited references from FDA Approval Summaries, direct PubMed searches, grants and peer review data. In the Williams study, assembly and analysis of each network took roughly 17 hours per drug, 10 of which are manual processing steps. While our process also involved expert level curation of a foundational set of references for each drug with a cost of roughly 2–5 hours, our network calculations and metrics ran in the order of minutes once the bibliometric data were assembled. Network visualization was performed using Cytoscape [18].

3. Results and discussion

Publications. Scientific publications form the backbone of each of these five networks. Our initial assumptions of appropriate citation and retrospective citation discovery (Introduction) suggest that network nodes that are common to multiple networks are likely to be influential. We calculated intersection counts for all possible combinations of publications in the Alemtuzumab, Imatinib, Nelarabine, Ramucirumab, and Sunitinib networks (Table 3). We also applied intersection analysis at a finer level of granularity by computing intersection counts for both first generation citations (citing_pmid) and second generation citations (cited_sid). The results are displayed as Venn diagrams in Figure 1.

Table 3. Intersection of five networks. Publications at the intersection of all five networks are listed above. All 14 publications are found in the second generation of references (cited_sid, Figure 1 right panel).

SourceYear	SourceName	Author(s)
1958	J. Am. Stat. Assoc.	Kaplan ER, Meier P.
1963	Science	Jerne, NK and Nordin, AA.
1972	J R Stat Soc	Cox DR.
1976	Anal. Biochem.	Bradford MM.
1977	Br J Cancer	R. Peto, M.C. Pike, and P. Armitage
1977	Proc. Natl. Acad. Sci.	Sanger FS., Nicklen S, Coulson AR.
1983	J Immunol Methods	Mosmann T.
1984	Adv Enzyme Regul	Chou TC, Talalay P.
1989	Molecular Cloning: A Laboratory Manual	Sambrook, J., Fritsch, E. and Maniatis, T.
1994	Acta Crystallogr D	Collaborative Computational Project 4
1994	Acta Crystallogr. A	Navaza J.
1997	Cell	Levine AJ.
1997	Am. J. Pathol.	Perez-Atayde AR, Sallan SE, Tedrow U, Connors S, Allred E, Folkman J.
1998	CA: A Cancer Journal for Clinicians	Landis SH, Murray T, Bolden S, Wingo PA.

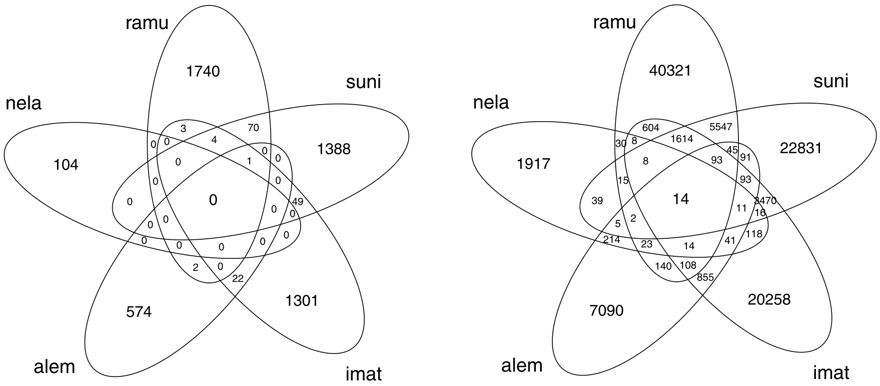

Figure 1. Intersecting publications in five networks. Intersections were calculated across all five networks for the first generation of references (citing_pmids) and as well as for the second generation of references (cited_sids) and displayed as Venn diagrams. *Left panel.* No first generation publications are observed common to all five networks. A single publication is cited in four of five networks. *Right panel.* 14 publications are common to all five networks. Abbreviations: alem (Alemtuzumab), imat (Imatinib), nela (Nelarabine), ramu (Ramucirumab), suni (Sunitinib).

The intersection of all five networks consists of 14 publications out of a total of 106,720 unique Scopus identifiers. Strikingly, not even a single publication is common to all five networks at the first generation level (citing_sid) although a single publication, the pathbreaking work of Kohler and Milstein on the production of monoclonal antibodies [19], is cited in four out of five networks. All 14 publications are in the second generation of citations (cited_sid) and another 198 comprise the sum of intersections in all possible four-network combinations, roughly an order of magnitude greater than the case of cited references. We manually grouped these 14 publications using high level descriptive terms and observed that this group was composed of statistical methods (5 publications), molecular and cell

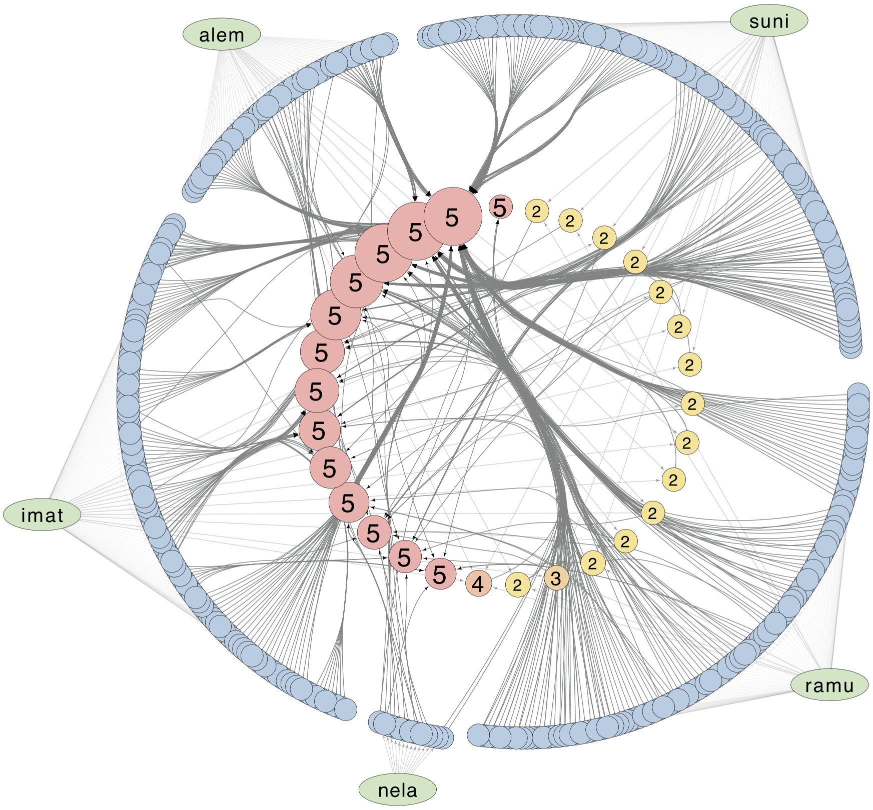

Figure 2. Core publications in networks. The outer arcs of blue nodes identify first generation publications (citing_sid) for each therapeutic. Nodes in the inner ring are sized by a gradient proportion to total degree count with an upper limit of 30 and are colored by a gradient proportional to the number of drug connections (2 to 5). 14 publications are common to all five networks (Table 3) and are colored red. The remaining nodes in the inner ring connect to between 2 and 4 drugs each and are labeled accordingly. Abbreviations: alem (Alemtuzumab), imat (Imatinib), nela (Nelarabine), ramu (Ramucirumab), suni (Sunitinib).

biological methods (4 publications), analytical and structural biology techniques (3 publications) and cancer biology (2 publications). Of these last two, one is a review of the p53 gene [20] and the second is a study of angiogenesis in children with acute lymphoblastic leukemia [21]. Thus, the majority of this small set of 14 publications describes methods that are heavily cited in these therapeutic development networks, which is consistent with observations of the general scientific literature [22]. Further, they support the concept of basic research contributing to subsequent innovation [23]. The relationship between core publications and their therapeutic networks is visualized in Figure 2. As the subject of another study, we are actively working on a scalable automated strategy to characterize the entire dataset as well as all combinations of intersections between networks using high level descriptive terms.

Grant support. With its annual budget of approximately US$32 billion, NIH is a major funder of biomedical research through its granting programs. Understanding

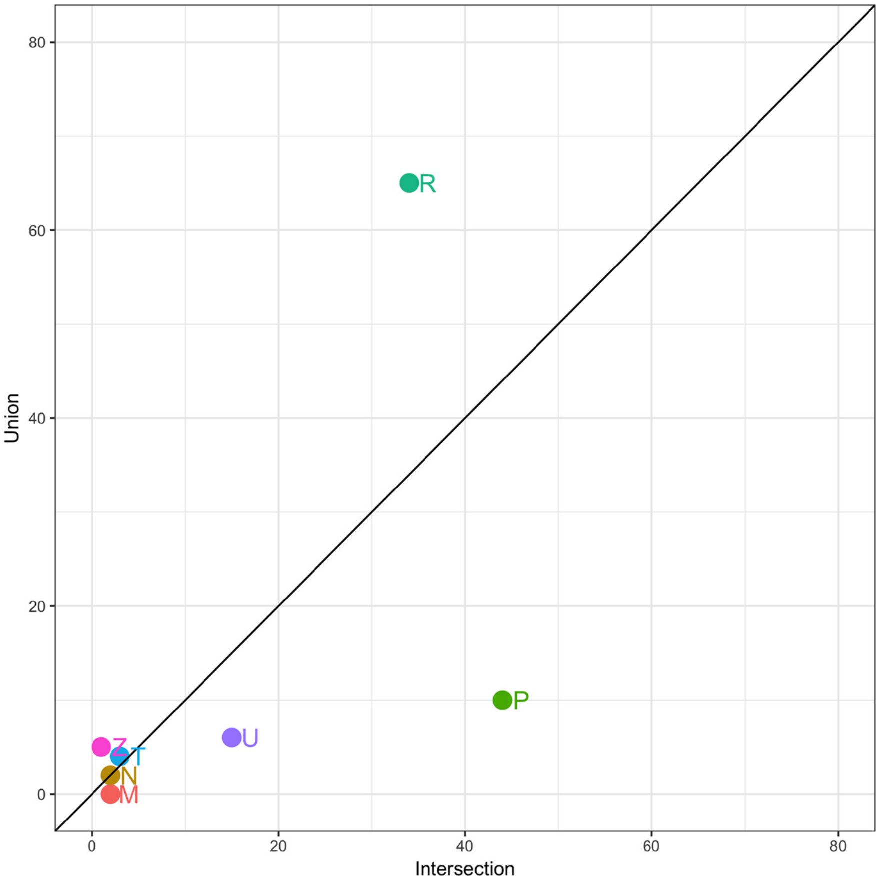

Figure 3. NIH research support. Grant and contract support for publications from NIH in the five networks was identified using ExPORTER data (Materials and Methods). 19,104 unique project numbers were identified as sources of support for publications in all five networks. Of these, 112 projects were common to all five networks. Projects were grouped by mechanism (i) P – Research Program Projects and Centers (ii) R – Research Projects (iii) M – General Clinical Research Centers Programs (iv) N – Research and Development-Related Contracts (v) U – Cooperative Agreements (vi) T – Training Programs (vii) Z – Intramural Research. For each mechanism, the number of projects in the intersection of all five networks was plotted against the number in the union of all five networks (both expressed as percentages of their respective totals). A higher proportion of Research Program Projects and Centers awards is found in the intersection group.

the nature and extent of NIH grant support for the research represented in our five networks, provides insight into the funding programs that enabled this research. We took advantage of publicly available data [13] to identify grant support for the publications in our five networks by mapping them to pmids. A total 19,104 unique grant numbers was harvested of which 112 were found in all five networks. At the intersection of five networks, the reason the number of grants is larger than the number of publications is because publications and grants exist in a ‘many to many’ relationship in that each publication can acknowledge support from multiple grants and each grant can support multiple publications. These awards were grouped by major type and visualized (Figure 3). Of note, support from Research Program

Projects and Center grants is proportionately larger in the intersection group when compared to the total population where the proportion of research projects is larger. A significant loss of information occurs when mapping from cited_sid to cited_pmid (Table 2). Thus we believe that these numbers may be an underestimate of actual grant support from NIH. Also missing from this analysis are details of research support from other funding agencies and industry, which are questions that we intend to pursue. Even so, these data testify to a recurring theme of collaboration and breadth of community engagement also seen at the publication level. We speculate that the broader and collaborative nature of such awards may be more likely to result in a methods-rich population of publications than the more focused research project award but elucidation will require further and more rigorous study.

Peer review. Research support from NIH is typically made through a two stage peer-review process. The Center for Scientific Review at NIH manages first-stage scientific review of between 50,000–60,000 grant applications each year [24], a process involving more than 15,000 expert reviewers. In addition, individual Institutes and Centers at NIH manage smaller scale peer review operations. Considering a crude estimate of a 20% success rate in funding, peer review can be viewed as a collaborative scientific activity and that serves as a selection layer for the upper fifth of applications, thus strongly influencing granting outcomes. To describe this layer at a high-level, we matched the awards in the five networks to the peer review panels (study sections) that evaluated them for scientific merit and calculated the intersection and union of these peer review panels. Eighty eight unique panel identifiers formed the intersection. Of these, 11 are distinguishable as Special Emphasis Panels that could be either one-time or recurring panels with temporary members, the remaining 77 are chartered panels with relatively stable membership. Some of these panels are no longer active and public records are not easily available to determine their scientific focus. For the 74 panels that could be classified, beyond an obvious focus on cancer, it is evident that the panels represent a rich mix of disciplines such as chemistry, biophysics, genetics, cell biology, and molecular biology; as well as AIDS, pathology, radiology, endocrinology, neurology, mental health, and child health. Four hundred and seven unique panel identifiers formed the union of all five networks. Of these 28 were Special Emphasis Panels, the remaining 379 panels were chartered as in the case of the intersection. These data provide evidence of broad input from invited experts in a collaborative activity that selects promising scientific projects. Assuming an average of 25 reviewers per panel (the number is likely to range from 5–40) and excluding that some of these panels are likely to have met multiple times during the lifespan of the awards in question and that some of these applications for funding may have been reviewed multiple times, a minimum of 10,000 experts comprised this additional layer of scientific influence. We believe that the actual number is likely to be at least double. A more accurate estimate would be possible if historical records of participation in peer review were made publicly available by NIH. We do not have records of funding awards or peer review conducted by organizations other than NIH and this is a focus of future investigation.

Network metrics. We have built upon the work of Williams [1] by addressing author disambiguation through the use of Scopus authorIDs and enriching the network data with grants and peer review data. Whereas, the original code was designed to handle a single drug and was not applied to a problem larger than 5000 articles, our approach scales to over 100,000 articles and our metrics include cross-network calculations. To quantify network data and to identify influential researchers in and across networks, we calculated PIR and RBR scores for all researchers as well as the new nPIR, PPR, nRBR, and gRBR scores (Materials and Methods). The nPIR metric describes the sum of aggregated citation scores for all articles that a researcher has in all five networks. The PIRpartitionRatio (PPR) normalizes the nPIR metric for an researcher to the sum of the researcher's individual PIR scores for each network being studied. A limitation of the nPIR and nRBR measurements is that they are valid only for the network(s) being studied. Scaling from five to the more than 1400 drugs approved by the FDA (and their many variants) would address this limitation [25], although other data related issues may well emerge.

While theoretically appealing (Materials and Methods), the gRBR is the most sensitive to data quality since it relies on an accurate estimate of total productivity of a researcher, which in turn depends on data quality in bibliometric databases. We found several instances in the top 10% of PIR scores where the gRBR was implausible likely on account of polysemy, synonymy, or incompleteness. This metric is therefore likely to be useful when article capture and the author disambiguation problem are resolved to the point where data quality is significantly improved and is not recommended except when strong confidence exists in the total productivity counts. These metrics may be best used in conjunction with positional measures such as quantiles to define populations of researchers within related networks, e.g., the top 25 researchers based on nPIR scores of all researchers in our dataset (Table 4). These 'bright stars' are elite performers in network(s) of clinical and basic science researchers that reinforce the concept of collaborative translational achievements built upon a body of basic science. Beyond simple aggregation, weighting, and normalization that we have used, a variety of citation metrics such as SNIP [26], with different normalization strategies at the field, journal, and article level are available for impact analysis and these could be applied to such networks depending on the features of these networks and the aim of the study [27]. The use of these citation measures will assume greater importance with scale up from small numbers of networks to a greater proportion of the global research network.

In this study, we begin with the literature directly cited in the approval process for five therapeutics. The approval process for therapeutics is subject to multiple layers of

Table 4. Elite performers. Researchers with the highest nPIR scores in the network of 5 anti-cancer therapeutics are listed. Also shown for each researcher is their PIRpartitionRatio (PPR). The nPIR indicates influence across all five networks and the PPR provides an estimate of how this influence is partitioned across each of the five networks (Materials and Methods). This list should be considered in the context of the data being analyzed and not interpreted as an absolute ordering of research excellence in the field.

Name	nPIR	PPR
Ferrara N.	46693	1.12
Folkman J.	23660	1.15
Ullrich A.	23034	1.46
Jain R.	15267	1.13
Heldin C.	15148	1.21
Druker B.	15088	1.16
Schlessinger J.	14996	1.48
Dvorak H.	14230	1.13
Alitalo K.	13812	1.24
Slamon D.	13587	1.46
Baselga J.	12908	1.36
Kantarjian H.	12029	1.17
Hicklin D.	11775	1.17
Witte O.	11449	1.08
Hanahan D.	11072	1.19
Buchdunger E.	11032	1.22
Risau W.	10950	1.24
Talpaz M.	10713	1.13
Mendelsohn J.	10534	1.54
Lydon N.	9988	1.18
Goldman J.	9927	1.11
Shibuya M.	9639	1.21
Kitamura Y.	9486	1.24
Waldmann H.	9363	1.04
Kerbel R.	9266	1.16

independent assessment, i.e., patent awards, clinical trials, and FDA reviews, each of which cite relevant literature. However, the only requirement of the current method is a starting set of articles. The method is applicable to any other therapeutic or set of therapeutics, but also to medical devices, diagnostics, standards development, and research discoveries and protocols such as induced pluripotent stem (iPS) cells [28] or CRISPR-Cas9 [29]. Thus, any question premised upon accessible citation records and defined by a nucleus of articles can be addressed with our approach. Furthermore, the method is also generalizable to sets of articles defined by groups of authors (e.g., consortia, programs), by institutions or academic departments, or by grant portfolios. The application of our method to the bibliographic record of an academic department, for example, would provide novel, network-based metrics by which to assess collaboration, productivity, and translational impact. Overall, no single metric will provide very useful answers, instead expert interpretation of multiple metrics best matched to curated datasets will more likely offer value. The method is implicitly time-dependent, allowing one to make historical comparisons and to set future goals. Importantly, our improvements to the method as presented here allow for the analysis across sets of articles in addition to within, supporting higher level interrogations of the crosstalk and collaborations that connect groups.

In summary, we have demonstrated a digital methodology based on data mining and network analysis, not restricted to drug discovery and cures alone, that offers burden-reduction in explorations of science history. The results argue that fundamental research, especially methods, found extensive application in these collaboration networks underlying the development of these five therapeutics. Thus, knowledge from basic research diffused over time to specific applications through the fields of biology, medicine, and pharmaceuticals. Beyond assembling a set of facts about a major scientific advance, the data assembled contribute to the understanding of collaboration across domains and can be used to enrich portfolio analysis, planning and optimization, as well as communications of the societal value of research. For the portfolio manager, such data, when coupled with aggregate measures, enable review on a scale that manual assembly would not permit. For the funder, an understanding of progress towards goals is supported. For the purpose of communication, these data on drug development provide the basis to explain that the work of thousands of scientists working in basic research benefits society since new drugs to combat disease have resulted from publicly supported research. An avenue for further work is application to historiography [30]. The data collected in this approach is largely time-stamped and the 'lag between non-mission research and the eventual innovations' [23] can be studied with respect to the discoveries common to multiple networks. The approach can be adapted to study the collaborative history within and across research portfolios of groups of researchers and targeted programs. While finer critical evaluation of the content of datasets generated through this approach and attempts to optimize resource allocation is best left to experts, the methodology is broadly accessible and can also be viewed as another tool for citizen science.

Declarations

Author contribution statement

George Chacko, Samet Keserci: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Lingtian Wan: Performed the experiments.

https://doi.org/10.1016/j.heliyon.2017.e00442

Eric Livingston, Alexander Pico: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare the following conflict of interests: Eric Livingston is an employee of Elsevier Inc., the publisher of the Scopus database. Samet Keserci, and George Chacko are employed by NET ESOLUTIONS Corporation. Lingtian Wan, was employed by NET ESOLUTIONS Corporation during the course of the study.

Additional information

Data associated with this study has been deposited at Mendeley Data under the accession number DOI: https://doi.org/10.17632/ysh53v7gpz.4.

Acknowledgments

This study would not have been possible without data that is made publicly available by NIH and the FDA. We thank Tandy Warnow from the University of Illinois Urbana-Champaign for critical comments and advice on formalizing definitions. We also thank Sandeep Somaiya and Syd Gomes from NETE as well as Holly J. Falk-Krzesinski, Daniel Calto, M'hamed Aisati, and Sherif El Shamy from Elsevier for their support of this collaboration.

References

[1] R.S. Williams, S. Lotia, A.K. Holloway, A.R. Pico, From scientific discovery to cures: bright stars within a galaxy, Cell 163 (2015) 21–23.

[2] C. Chen, D. Hicks, Tracing knowledge diffusion, Scientometrics 59 (2004) 199–211.

[3] J. Maldame, The importance of the history of science in intellectual formation, in: Scripta Varia, vol. 104, 2002, pp. 237–248.

[4] M.S. Lauer, PCSK9 inhibitors: lots of work done, lots more to do, Ann. Intern. Med. 164 (9) (2016) 624–625.

[5] J.J. O'Shea, Y. Kanno, A.C. Chan, In search of magic bullets: the golden age of immunotherapeutics, Cell 157 (2014) 227–240.

[6] E. Buchdunger, J. Zimmermann, H. Mett, T. Meyer, M. Müller, B.J. Druker, N.B. Lydon, Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative, Cancer Res. 56 (1) (1996) 100–104.

[7] A.M. O'Farrell, T.J. Abrams, H.A. Yuen, T.J. Ngai, S.G. Louie, K.W. Yee, et al., SU11248 is a novel FLT3 tyrosine kinase inhibitor with potent activity in vitro and in vivo, Blood 101 (9) (2003) 3597–3605.

[8] P.T. Ho, B.D. Cheson, P.H. Phillips, Clinical trials referral resource. Clinical trials using compound 506U78, Oncology (Williston Park) 10 (12) (1996) 1831–1832.

[9] Y. Krupitskaya, H.A. Wakelee, Ramucirumab, a fully human mAb to the transmembrane signaling tyrosine kinase VEGFR-2 for the potential treatment of cancer, Curr. Opin. Investig. Drugs 10 (6) (2009) 597–605.

[10] G. Hale, S. Bright, G. Chumbley, T. Hoang, D. Metcalf, A.J. Munro, H. Waldmann, Removal of T cells from bone marrow for transplantation: a monoclonal antilymphocyte antibody that fixes human complement, Blood 62 (4) (1983) 873–882.

[11] Federal Drug Administration, Drugs@FDA: FDA approved drug products, https://www.accessdata.fda.gov/scripts/cder/daf/.

[12] Google Inc., Google patents, https://patents.google.com/.

[13] National Institutes of Health, https://exporter.nih.gov/.

[14] S. Keserci, E. Livingston, Lingtian Wan, A. Pico, G. Chacko, Bright stars in neighboring constellations v4, Mendeley Data.

[15] S. Keserci, E. Livingston, L. Wan, A.R. Pico, G. Chacko, Nete Labs case studies, https://github.com/NETESOLUTIONS/NETELabs_CaseStudies.

[16] S. Lotia, A.P. Pico, Gladstone bibliometrics, https://github.com/gladstone-institutes/bibliometrics/tree/1.0.1.

[17] M.E. Falagas, E.I. Pitsouni, G.A. Malietzis, G. Pappas, Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses, FASEB J. 22 (2) (2008) 338–342.

[18] P. Shannon, A. Markiel, O. Ozier, N. Baliga, J.T. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.

[19] G. Köhler, C. Milstein, Continuous cultures of fused cells secreting antibody of predefined specificity, Nature 256 (1975) 495–497.

[20] A.J. Levine, p53, the cellular gatekeeper for growth and division, Cell 88 (3) (1997) 323–331.

[21] A.R. Perez-Atayde, S.E. Sallan, U. Tedrow, S. Connors, E. Allred, J. Folkman, Spectrum of tumor angiogenesis in the bone marrow of children with acute lymphoblastic leukemia, Am. J. Pathol. 150 (3) (1997) 815–821.

[22] R. Van Noorden, B. Maher, R. Nuzzo, The top 100 papers, Nature 514 (2014) 550–553.

[23] F. Narin, Tracing the paths from basic to economic impact, F&M Sci. (2013) 67–87.

[24] K.W. Boyack, M.-C. Chen, G. Chacko, Characterization of the peer review network at the Center for Scientific Review, National Institutes of Health, PLoS ONE (2014).

[25] M.S. Kinch, A. Haynesworth, S.L. Kinch, D. Hoyer, An overview of FDA-approved new molecular entities: 1827–2013, Drug Discov. Today 19 (8) (2014) 1033–1039.

[26] L. Waltman, N.J. van Eck, T.N. van Leeuwen, M.S. Visser, Some modifications to the SNIP journal impact indicator, J. Informetr. 7 (2) (2013) 272–285.

[27] J.P.A. Ioannidis, K. Boyack, P.F. Wouters, Citation metrics: a primer on how (not) to normalize, PLoS Biol. (2016).

[28] K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors, Cell 126 (4) (2006) 663–676.

[29] E. Charpentier, J.A. Doudna, Biotechnology: rewriting a genome, Nature 495 (7439) (2013) 50–51.

[30] E. Garfield, Citation indexes for science: a new dimension in documentation through association of ideas, Science 122 (1955) 108–111.

https://doi.org/10.1016/j.heliyon.2017.e00442