Published in final edited form as:

Front Res Metr Anal. 2017 May ; 2: . doi:10.3389/frma.2017.00003.

Gaps within the Biomedical Literature: Initial Characterization and Assessment of Strategies for Discovery

Yufang Peng[1], **Gary Bonifield**[2], and **Neil R. Smalheiser**[2,*]

[1]School of Information Management, Nanjing University, Nanjing, China

[2]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 60612 USA

Abstract

Within well-established fields of biomedical science, we identify "gaps", topical areas of investigation that might be expected to occur but are missing. We define a field by carrying out a topical PubMed query, and analyze Medical Subject Headings by which the set of retrieved articles are indexed. Medical Subject headings (MeSH terms) which occur in >1% of the articles are examined pairwise to see how often they are predicted to co-occur within individual articles (assuming that they are independent of each other). A pair of MeSH terms that are predicted to co-occur in at least 10 articles, yet are not observed to co-occur in any article, are "gaps" and were studied further in a corpus of 10 disease-related article sets and 10 related to biological processes. Overall, articles that filled gaps were cited more heavily than non-gap-filling articles and were 61% more likely to be published in multidisciplinary high-impact journals. Nine different features of these "gaps" were characterized and tested to learn which, if any, correlate with the appearance of one or more articles containing both MeSH terms within the next five years. Several different types of gaps were identified, each having distinct combinations of predictive features: a) those arising as a byproduct of MeSH indexing rules; b) those having little biological meaning; c) those representing "low hanging fruit" for immediate exploitation; and d) those representing gaps across disciplines or sub-disciplines that do not talk to each other or work together. We have built a free, open tool called "Mine the Gap!" that identifies and characterizes the "gaps" for any PubMed query, which can be accessed via the Anne O'Tate value-added PubMed search interface (http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi).

Keywords

Scientific discovery; link prediction; Medical Subject Headings; text mining; literature based discovery

[*]**Correspondence:** Neil R. Smalheiser, neils@uic.edu.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Conceived and designed the study (NS), provided data on PMIDs and feature scores (GB), programmed the website (GB), calculated network features and citations (YP), carried out statistical analyses (NS, GB), performed regression analyses (NS), and wrote and approved the manuscript (NS, YP, GB).

Introduction

Although no single theoretical framework covers all types of scientific discoveries, several scholars have pointed out that new discoveries often involve new combinations of existing concepts or ideas (Swanson and Smalheiser, 1997; Chen et al, 2009; Uzzi et al, 2013; reviewed in Chen, 2013). For example, published articles often present pre-existing text terms (Packalen and Bhattacharya, 2015), Medical Subject Headings (MeSH terms) (Theodosiou et al, 2011; Mishra and Torvik, 2015) or citations to journals (Uzzi et al, 2013) that are combined in novel ways not seen before. It is not clear whether articles that combine novel pairs of MeSH terms will necessarily result in more impactful or important research as assessed by (say) patents or citations. However, pairs of MeSH terms have a certain appeal as an objective measure of information flow, since one can examine the overall number of new MeSH pairs that appear over time within a given field of investigation, as well as attempt to predict which new MeSH pairs are most likely to appear.

In previous investigations, we have studied the potential benefit of identifying disparate areas of scientific investigation which are disconnected – that is, they reside in two different sets of articles that do not share authors, and are poorly cross-cited or co-cited – yet they contain information that, when connected, leads to promising and testable new hypotheses (Swanson and Smalheiser, 1997; Torvik and Smalheiser, 2007; Smalheiser et al, 2009). The presumption is that connections between disparate areas of investigation are likely to be overlooked (due to lack of reading widely enough by scientists), neglected (e.g. because they do not correspond to existing mainstream topics), or disfavored as implausible or meaningless (e.g., one might not expect much insight to emerge from connecting studies of in vitro fertilization and handgun safety) (Swanson, 1986).

Here, we consider the natural history of un-connected topics of investigation that reside WITHIN a single well-established field of study. MeSH term pairs that have expected co-occurrences of 10 or more within a given field, but do not co-occur in any articles in that field, are defined as “gaps”. That is, we seek to bridge a set of articles that are indexed by MeSH term 1, and those that are indexed by MeSH term 2, all within the larger set of articles represented by a topical PubMed query. Such gaps are less likely to represent neglected or overlooked relationships, and in fact, some of the un-connected topics might represent low-hanging fruit that investigators will explore with high priority in the near future.

In the present paper, we have identified gaps from a variety of topical PubMed queries, comprising 10 disease-related article sets and 10 related to biological processes. The gaps have been characterized in terms of 9 different features at one initial time window (1987–2005). We then identified articles appearing in the same field in the subsequent five year time window (2006–2010), and looked at those that did vs. did not fill one or more gaps (that is, articles dual-indexed with both MeSH terms of the gap). We hoped to learn how gap-filling articles differ from those that do not fill gaps. Moreover, comparing different gaps among each other, we assessed whether we can identify features which correlate with the likelihood that the gap will be filled in the second time window, and with the number of articles that fill that gap.

Materials and Methods

A total of 20 PubMed queries was carried out using the PubMed eUtils API, representing 10 diseases and 10 biological processes (Table 1). These were chosen to cover a wide range of topical areas, pathologies and tissue systems, and were divided into two time slices. The first time slice consisted of articles with publication dates 1987–2005 inclusive (we did not include articles earlier than 1987 because the number and indexing of Medical Subject Headings (MeSH terms) has evolved substantially over time). The second time slice consisted of articles with publication dates 2006–2010 inclusive, i.e., the five year period following the first time slice. As shown in Table 1, the number of articles in the first time slice ranged between 7,000 and 50,000, and the new articles appearing in the second time slice increased the overall size of each literature by 1.25-fold to 2-fold.

For each set of articles retrieved from each PubMed query (which will be referred to as the query set, the retrieval set, or the topical literature), in the first time slice, we extracted all MeSH terms (ignoring subheadings) and all pairs of MeSH terms co-occurring in the same article. For MeSH terms that occurred in at least 1% of the articles in the query, we computed the co-occurrence of MeSH term pairs that would be expected if each MeSH term is assigned to articles independently, at random. Those MeSH term pairs that had expected co-occurrences of 10 or more (i.e., Frq (query AND MeSH1) * Frq (query AND MeSH2) / number of articles ≥ 10), but an observed co-occurrence of 0, were defined as “gaps”. (Note that although articles published before 1987 were not included in the first time slice, no MeSH term pair was counted as a gap if one or more articles published before 1987 were indexed by both of the MeSH terms. Thus, gaps reflect the entire MEDLINE literature from its beginnings through the end of 2005.)

Counting citations

To compute citations for articles published in the second time slice, we used the title, publication date and DOI when available to identify the Google Scholar (GS) record and to extract the listed Google Scholar citations as of November 2016. Of 2,418 gap-filling articles, 6 were not found in Google Scholar; for those, we used their citations in PubMed Central instead.

Within each query, each gap was characterized according to different features:

1. The expected number of co-occurrences within the query set;
2. The expected number of co-occurrences within MEDLINE as a whole (using the frequencies of each MeSH term within all of MEDLINE in the baseline 2015 release);
3. The observed number of co-occurrences within MEDLINE as a whole (baseline 2015);
4. The article odds ratio (= observed / expected number of co-occurrences within MEDLINE as a whole; Smalheiser and Bonifield, 2016).

5. The author odds ratio. This is computed as the observed / expected number of co-occurrences within the body of articles written by an individual author, summed over all authors publishing in MEDLINE (Smalheiser and Bonifield, 2016).

6. The pR score. This is an innovative measure of semantic similarity between the articles within the query set indexed by the first MeSH term vs the articles indexed by the second MeSH term (Torvik and Smalheiser, 2007). The Arrowsmith two-node search interface (http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/start.cgi; Smalheiser et al, 2009) is utilized to carry out two separate PubMed searches, [query AND MeSH term 1] vs. [query AND MeSH term 2]. In both cases, the MeSH terms are searched without expansion to retrieve related terms. Arrowsmith software computes the words and phrases that are in common to the titles of articles in the two queries (ie., the B-terms), and uses a quantitative model (Torvik and Smalheiser, 2007) to estimate the predicted relevance of each B-term for linking the two queries in a meaningful way. The percentage of B-terms that are predicted to be relevant is the pR score. We have hypothesized that the pR score may provide a metric to measure the overall implicit information shared by two sets of articles defined by PubMed topical queries (Torvik and Smalheiser, 2007). Because the two-node search only gives meaningful results when the two queries are of sufficient size, we only calculated pR scores when the geometric mean of Frq (query AND MeSH1) and Frq (query AND MeSH2) ≥ 2% of the total number of articles in the query.

7. The Common Neighbors (CN) score. This is calculated for each query by making a network graph of all MeSH terms that occurred in at least 1% of the articles in the query, where each MeSH term is a node, and nodes are linked if they co-occur in at least one article in the query. For each MeSH term pair that represents a gap, we calculate the number of common neighbors. Python2.7 and the Networkx-1.10 framework package were used for this and the next feature.

8. The Adamic-Adar (AA) score (Adamic and Adar, 2003). This is a normalized CN score in which the contribution of each common neighbor is divided by the log number of links that it has. The formula is SAA(u,v)= $\sum_{w \in «ŭ(u) \cap T(v)} 1/log\,|«ŭ\,(w)|$ where $«ŭ\,(u)$ denotes the set of neighbors of u.

9. The MeSHSim score. This is a measure of similarity between the two MeSH terms according to path distance on the MeSH hierarchy. We used the R package by Zhou et al, 2015, using headingSim with parameters headingSim (mesh1, mesh2, method="SP", frame="node").

The outcomes analyzed for each gap in this study were:

1. Number of articles appearing in the second time slice which were indexed by both MeSH terms comprising the gap. We refer to such articles as those which fill the gap.

2. Presence or absence of articles appearing in the second time slice which fill the gap, scored 0 or 1, regardless of the exact number of articles.

3. Citations per gap (CPG). For each gap, we computed the square root of the arithmetic mean of the number of citations across all articles filling that gap. (Since citations follow a power law approximately, the square root was taken to make the data distribution more quasi-normal.)

4. Maximum citation per month (MCM). For each gap, we identified the article having the maximum number of citations, and normalized that to citations per month, by dividing by the number of months from its publication date to November 2016.

Note that in a few cases, the same gap appeared in more than one query, and was processed separately (since some of the feature scores are query-specific).

For each query, to compare gap-filling vs. non-gap-filling articles, the PMIDs corresponding to each set of articles were entered into our value-added PubMed search interface Anne O'Tate (http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi; Smalheiser et al, 2008). We extracted the top 20 journals, MeSH terms, and author names for the articles that did, vs. did not, fill at least one gap. We also tabulated the percentage of articles that were published in a multidisciplinary high-impact journal (defined as either *Science, Nature, Proc. Natl. Acad. Sci. USA, Cell, New England Journal of Medicine, the Lancet*, or *JAMA*). Finally, to compare citations of gap-filling vs. non-gap-filling articles, for each of the 20 queries, we randomly chose non-gap-filling articles (the number to equal the gap-filling articles for that query). That is, an equal number of gap-filling and non-gap-filling articles were compared for each query.

Results

Prevalence and fate of gaps across queries

Gaps – defined here as pairs of MeSH terms that never co-occurred within any article in the query set despite an expected co-occurrence of 10 or more based on their individual frequencies – were surprisingly prevalent within the PubMed query sets. The number of gaps ranged widely from 10 to 889 across different queries (Table 2). This variability was only partially accounted for by differences in the sizes of the query sets, since queries ranged from 0.13 to 2.78 gaps per 100 articles (Table 2). Nonetheless, the mean number of gaps per query was 261, representing a great number of potential "loose ends" that one might potentially attempt to tie up in the near future!

In fact, across all queries, 30.7% of the gaps present at the end of 2005 were filled by one or more articles published during the subsequent five year period of 2006–2010. The high rate at which gaps were filled suggests that they do not generally represent neglected back-waters or research dead-ends. We compared articles that do vs. do not fill gaps in the second time slice, to see which types of journals they were published in. We found that the vast majority of gap-filling articles appeared in journals sharing the topical scope of the top 20 journals where non-gap-filling articles appeared, and only about 10% of the journals were topically

divergent. For example, for the acute myocardial infarction query, 27 of the 30 articles that filled gaps were published in cardiology journals; plus one published in a psychiatry journal, one in an oncology journal, and one in a toxicology journal. Similarly, on average, the top 20 MeSH terms indexing the gap-filling articles shared 12 MeSH terms with the top 20 MeSH terms that indexed the non-gap-filling articles. Moreover, there was some overlap of authors between the two sets: Across all queries, an average of 1.45 author names (listed on 2 or more gap-filling articles) were also in the top 20 author names in the list of those publishing non-gap-filling articles.

Impact of gap-filling articles

Gap-filling articles were more highly cited, overall, than non-gap-filling articles (65.11 vs. 52.44 citations on average), a difference that was highly significant and consistent across the dataset as a whole (nonparametric two-tailed Mann-Whitney U test, $p = 4.11 \times 10^{-14}$; unpaired two-tailed t-test performed on square roots of citations, $p = 7.82 \times 10^{-11}$). Thus, insofar as citations are indicative of scientific impact, gap-filling articles were more prominent than a random cross-section of articles published on the same topics.

Perhaps most tellingly, gap-filling articles were significantly more likely to appear in the most prestigious, multidisciplinary high-impact journals (defined here as including *Science, Nature, Proc. Natl. Acad. Sci. USA, Cell, New England Journal of Medicine, the Lancet*, and *JAMA*). Overall, 3.34% of gap-filling articles were published in high-impact journals compared to 2.08% of non-gap-filling articles (3.34 ± 3.19 SD vs. 2.08 ± 1.22, N = 20, p = 0.016 by paired two-tailed t-test; p = 0.0026 by sign test). Stated another way, gap-filling articles are 61% more likely to appear in high-impact journals than non-gap-filling articles (3.334/2.08 = 1. 606). The set of gap-filling articles is enriched in findings deemed particularly novel and significant at the time of publication (i.e., as assessed by reviewers and editors of high-impact journals). Together, these findings suggest that most of the gap-filling articles reside in, and may help redirect, the mainstream of the field.

Features of gaps in disease vs. biological processes query sets

The 10 disease query sets and 10 biological processes query sets were comparable in terms of number of articles per query. The number of gaps, and the percentage of gaps filled in the second time slice, were not significantly different between the two types of queries. However, the disease-related vs. biological processes article sets differed significantly in their structure. For example, the biological processes queries had almost twice the number of MeSH pairs per article as the disease queries (Table 2), suggesting that they are more topically diverse. The article odds ratios of the gaps found in the disease query sets were less than half the mean value in the biological processes queries (Table 3). The MeSH terms comprising disease gaps had significantly lower similarities as judged by author odds ratios, CN and AA network scores, and MeSHSim as well (Table 3). Interestingly, the distribution of pR scores was nearly the same in the two types of queries (mean = 0.252, SD = 0.082 vs. mean = 0.254, SD = 0.086; Table 3). Yet only 56% of the disease gaps were bridging MeSH terms that were frequent enough to calculate pR scores (see Methods), whereas 72% of biological processes gaps were scored. We had not initially anticipated that these parameters would differ across the board so strikingly, especially since both article sets are biomedical

in nature. This finding presumably reflects the fact that biological processes are studied across multiple model systems and multiple levels of integration (from molecules to organisms), whereas the disease-related studies are more narrowly focused on humans and issues such as diagnosis, pathogenesis and treatment. In any case, this led us to analyze the features of gaps within the article sets both separately and as a combined dataset.

Supplementary Table 1 shows the nonparametric Spearman rank correlation rho values for the features and outcomes all considered pairwise, for the combined dataset and separately for disease and biological processes queries. (Note that Pearson linear correlation values are not appropriate here since many of the features and outcomes are not normally distributed.) Considering, first, how different gap features are correlated among themselves, it is apparent that the CN and AA scores are almost perfectly correlated (0.994), suggesting that they are redundant, whereas the MeSHSim score shows a very low correlation with any other feature (<0.20), suggesting that that it measures a very different and non-redundant type of MeSH term similarity. The article odds and author odds show an intermediate correlation (0.5-0.6) as expected (Smalheiser and Bonifield, 2016), indicating that they are related but that each gives information on its own. The article odds measures how often the two MeSH terms co-occur in MEDLINE as a whole (not just within the given query), relative to what is expected by chance, whereas the author odds measures how often the two MeSH terms co-occur in the body of articles written by the same individual, relative to the level expected by chance.

Gaps are heterogeneous

We identified several types of gaps that arise from different scientific scenarios, each associated with a different combination of features:

1. **Gaps arising as a byproduct of MeSH indexing.** A lack of co-occurrence of two MeSH terms could potentially be a trivial consequence of MeSH indexing rules. For example, few articles are indexed with both “Fatal Outcome” and “Mortality”, despite the very similar nature of these two MeSH terms. This is because MEDLINE indexers are instructed to index an article with “Fatal Outcome” if it is concerned with death of an individual, whereas an article is indexed with “Mortality” if it discusses death at a population level. Another example is “Tumor Cells, Cultured” vs. “Cell Line, Tumor”. These are closely related; in fact, they are adjacent on the MeSH hierarchy and have a very high MeSHSim score = 0.92, but the former concept is used to index articles that describe culturing tumor cells acutely whereas the latter term is used if the tumor cells are established as a cell line. Across the combined dataset of 5221 gaps, only 3 gaps had MeSHSim scores above 0.9, and all were examples of MeSH indexing rules.

2. **Gaps that lack biological meaning.** Another reason that two MeSH terms might not co-occur within the same article is because there is truly no meaningful relationship between them. For example, separate articles on suicide discuss “Suicide, Assisted” and discuss the enzyme “Thymidine Kinase”, but arguably there is no information gain from combining these apparently unrelated concepts. Note that the article odds is very low for this gap (= 0), as is the author odds (=

0.063). The pR score for these two MeSH terms is only 0.045, similar to the value observed when pairs of MeSH terms are chosen at random (Torvik and Smalheiser, 2007). Gaps of this nature are much less likely to be filled during the second time slice: Across the combined dataset, 862 gaps satisfied the criteria (article odds <0.2, author odds <0.5, pR <0.2), of which only 9.3% were filled during the second time slice, with a mean of 1.27 articles per filled gap (compared to 28.2% across all other gaps with a mean of 1.89 articles per filled gap). Using stricter criteria (article odds <0.1, author odds <0.2, pR <0.2), there were 71 gaps of which only 2.8% were filled (by one article each).

3. **Gaps that represent "low hanging fruit".** Conversely, a gap may represent "low hanging fruit" – pairs of MeSH terms that have not previously been studied together yet both lie at the research frontier of the field. For example, "Angina, Unstable" and "Heart Rupture, Post-Infarction" did not co-occur in any article within the acute MI literature by the end of the first time slice, but these two topics did co-occur within Medline as a whole as often as expected by chance (i.e., article odds = 1.06), and they had an extremely high author odds score (= 36.42), indicating that the same investigators had a strong tendency to discuss both topics (albeit in different articles). Two articles in the second time slice were indexed by both of these MeSH terms, and the concepts were explicitly related to each other (i.e., heart rupture was a complication of unstable angina-associated infarction). Across the combined dataset, we observed 185 gaps satisfying the criteria (article odds >1, author odds >1), of which 47.6% were filled during the second time slice, with a mean of 3.4 articles per filled gap.

4. **Gaps in communication.** A final (and perhaps the most interesting) reason that gaps may exist in the first time slice is because the two MeSH terms are associated with different disciplines, groups of investigators, geographical regions, or other sub-groups of the field that are either not aware of each other, or do not collaborate together. For example, in the apoptosis query set, the pair of MeSH terms "Liver Neoplasms" and "Neurons" remained unfilled in the second time slice, and have very low article odds (= 0.0065) and low author odds (= 0.40). Nevertheless, since liver cells and neurons share many biochemical and cellular pathways, it is plausible that they share information that might be useful to bridge. Note that the pR score for this pair is 0.499, similar to the highest values observed for pairs of article sets that are closely topically related (Torvik and Smalheiser, 2007). Such a gap may be intrinsically promising, yet remain unexplored during the second time slice, due to reasons that may be cultural or pragmatic rather than scientific. Across the combined dataset, we observed 45 gaps satisfying the criteria (article odds <0.2, author odds <0.5, pR >0.35), of which 22.2% were filled during the second time slice, and a mean of 1.2 articles published per filled gap (compared to 25.1% filled across all other gaps, and 1.85 articles published per filled gap). Thus, this type of gap is not entirely neglected, but the number of articles published per gap is relatively low, especially compared to the "low hanging fruit". As indicated above, the influence of pR was restricted to the diseases queries: 36.4% of gaps fulfilling the criteria for this

type were filled during the second time slice, and a mean of 1.2 articles published per filled gap, in contrast to the biological processes queries where only 8.7% of these gaps were filled and 1.0 articles per filled gap.

Table 4 shows the gaps of this type within the biological processes queries. Most of these gaps are bridging very different disciplines (e.g., Infertility, Male::Saccharomyces cerevisiae) and –at first glance, at least – might be thought to share no interesting information, except that their high pR score points to interesting information that might possibly have been overlooked.

Features that correlate with the likelihood that a gap will be filled in the near future

Putting aside the heterogeneity of gaps, we next analyzed how individual gap features were predictive of gap-filling across the combined queries and in disease-related vs. biological processes queries. We were able to identify certain gap features that correlated with the likelihood that a given gap will be filled in the next five years, and with the number of articles filling the gap during that period. The best single predictor overall was the article odds, rho = 0.35 (0.38 in disease queries, 0.32 in biological queries). Stated another way, MeSH pairs which have co-occurred frequently elsewhere in MEDLINE are also the most likely to appear in the given query literature in the near future. This is likely to reflect gaps that are “low hanging fruit”. Other individual features that correlated with the subsequent number of gap-filling articles in the disease queries were pR (= 0.34); CN/AA (= 0.34); observed co-occurrence in Medline (= 0.32); and author odds (= 0.23). The rho correlations were generally lower in the biological processes queries (observed co-occurrence in Medline (= 0.27); CN/AA (= 0.24); author odds (= 0.10), pR (= 0.02)).

In order to see which features were most important after holding all other factors constant, and to see which combinations of features were most predictive across all gaps, we used the Weka 3.8.0 software environment (Frank et al, 2016) to explore multiple linear regression models. Features that had skewed distributions were converted to square root values to improve quasi-normality, and the software was set to adjust for feature scaling and collinearity. Cross-validation (10-fold) was applied to avoid overfitting. The goal was not to optimize model predictive performance, but to understand the relative weights and independent influences of each feature. As shown in Table 5, over the entire dataset of 20 queries, the pR score was the most important predictor of the number of gap-filling articles that will appear in the second time slice, followed by the article odds. The relative importance of features was the same in both groups for classifying gaps simply as filled or not filled in the second time slice (0 vs. 1), regardless of the number of articles (data not shown). Relative importance of features was also the same when gaps whose article odds = 0 were removed from the set (i.e., by definition, none of those can have publications in the second time slice; not shown). A random forest model with the same features produced about the same performance as the multiple linear regression model (not shown).

As suspected, the fitted regression models were quite different for the disease queries vs. the biological processes queries (Table 6 vs. Table 7). Interestingly, both article odds and pR scores were the top two predictors in both groups, although pR was far more important in the disease queries whereas article odds dominated in the biological processes queries. The

MeSHSim score had an appreciable negative weight in the disease queries only, whereas the author odds ratio had an appreciable negative weight only in the biological processes queries. The network similarity scores (Common Neighbors, CN, and Adamic-Adar, AA) contributed significantly when examined as single features in both sets of articles (Table 5, cf. Kastrin et al, 2016) but had little effect in the fitted models; that is, their influence could be "explained away" by the influence of other correlated features.

Relation of gap features to subsequent citations

Surprisingly, none of the features that were examined showed any large rank correlations with the number of citations garnered by gap-filling articles, either measured as an average of the articles filling a particular gap, or the paper having the maximum number of citations (Supplemental Table 1). This lack of correlation was observed for both disease-related queries and biological processes queries (Supplemental Table 1) and for the subset of "low hanging fruit" (not shown). Normalizing the number of citations by topic (i.e., dividing the citations for each gap by the mean number of citations observed for gaps in that query) did not affect these negative results. Thus, although gap-filling articles as a group tended to garner more citations than non-gap-filling articles, we were not able to identify features that predicted which gaps, once filled, were more likely to be highly cited. These findings are possibly limited by the fact that only one time slice and duration were examined, and some transformational research may only garner citations after substantial time has passed (e.g., van Raan, 2004).

Discussion

The present paper identified and characterized "gaps", i.e., pairs of MeSH terms that never co-occurred within any article in a given topical biomedical literature, despite an expected co-occurrence of 10 or more based on their individual frequencies. Ten disease queries and ten biological processes queries were conducted in PubMed to generate a test bed in the range of 7,000 – 50,000 articles each for study. We found that each query set (containing articles through 2005) is associated with multiple gaps. As a whole, these are relatively dynamic, since almost a third of them were filled within the next five years by the publication of one or more articles that were dual indexed with both MeSH terms. The gap-filling articles are published in journals that are topically similar to non-gap-filling articles, are more highly cited than non-gap-filling articles on the same topic, and are 61% more likely to be published in multidisciplinary high-impact journals.

Gaps fell into several categories that have very different origins, characteristic features, and implications for scientists. A total of 9 gap features were employed to characterize gaps, which were used to divide them into several categories:

a. Gaps arising as a byproduct of MeSH indexing rules that constrain dual indexing of very similar MeSH terms. The signature of these infrequently observed gaps was a very high MeSHSim score, which measures path closeness on the MeSH hierarchy.

b. Gaps that lack any obvious biological meaning. Such gaps had very low scores on multiple features.

c. Gaps that represent “low hanging fruit”. Such gaps had relatively high article odds scores, meaning that the pair of MeSH terms has previously been well studied in other fields and is now poised to contribute to the given query literature. This gap category appears to be most likely to generate gap-filling articles in the next five years.

d. Gaps in communication. These have not been brought together in any article because the two MeSH terms are associated with different disciplines, groups of investigators, geographical regions, or other sub-groups of the field that are either not aware of each other, or do not collaborate together. Such gaps had low article odds (that is, they had not been well studied together anywhere in MEDLINE) and low author odds scores (that is, the same investigators had rarely written on both topics, even in separate articles), yet such gaps showed evidence that the MeSH terms generate some useful information when brought together, as reflected by high implicit similarity pR scores (Torvik and Smalheiser, 2007). We have hypothesized that the pR score measures the amount of implicit information shared between two MeSH terms (Torvik and Smalheiser, 2007), and a high pR score ought to imply that significant new knowledge can be gained from bridging the two MeSH terms (Swanson and Smalheiser, 1997). Thus, in contrast to “low hanging fruit”, the “gaps in communication” may point to discoveries that involve more unexpected or surprising new connections.

Although this initial categorization of gap types appears to hold generally, the relative prevalence of different types of gaps varied extensively across biomedical article sets on different topics which are structured in different ways (e.g., some are topically narrow and focused, others are a mix of separate communities). Multiple linear regression modeling was carried out to learn which gap features, if any, correlated with the number of gap-filling articles appearing in the next five years. However, we learned that fitting a single quantitative model was of limited value, because of the heterogeneity in types of gaps and types of queries. For example, the pR score was the strongest predictor of gap-filling articles when tested in disease-related query sets, yet the article odds ratio was a much stronger predictor in biological processes query sets. When considering all gaps as a single collection, the features studied here explained only about 16% of the variability in the number of articles published in the second time slice (Table 5).

The initial study reported here provides the starting point for much further research. For example: What is the function and relative importance of articles that fill gaps – that is, articles newly combining MeSH terms that are both already well represented in a field – relative to articles that introduce MeSH terms appearing in the disciplinary field for the first time, and relative to articles that are indexed by MeSH terms which are entirely novel and newly added to the MeSH hierarchy?

- How important are different types of gaps, particularly “low hanging fruit” vs. “gaps in communication”, for driving the mainstream of a field in new directions? Which types of gaps, if any, are most likely to lead to findings that transform a field radically?

- Which types of gaps, once filled, will be associated with articles that are cited highly? None of the gap features that we examined were useful in predicting the number of citations that a gap-filling article will garner, but we only examined the set of gaps as a whole. This analysis needs to be repeated using a larger dataset of queries that examines different types of gaps separately.
- As well, when comparing gap-filling vs. non-gap-filling articles on the same topic, it will be interesting to see how they compare on other features that reflect research strategies and approaches, such as journal, country of origin, size of collaborative author team, extent of interdisciplinarity (Larivière et al, 2015), and so on.
- The present study only examined one five year follow-up time slice (2005–2010). However, that time frame was chosen rather pragmatically. It will be worth examining gaps filled during the next five years (2010–2015) to see if the results are similar. Also, the time of delay in filling a gap is itself a feature that may have significance.
- What is the significance of gaps that have never co-occurred in any articles in MEDLINE to date, i.e., those whose article odds = 0? Filling such gaps requires truly novel combinations of MeSH terms. Although most of these gaps may not be biologically meaningful, those that do bridge useful information may have the potential to be especially surprising and innovative.
- Finally, there are inherent limitations to the use of MeSH terms for identifying gaps, due to their manual assignment to articles, partial redundancy in some cases, granularity, incomplete coverage, relatively slow updating of new terms into the hierarchy, restriction to articles indexed in MEDLINE, and so on. The Arrowsmith project (Torvik and Smalheiser, 2007; Smalheiser et al, 2009) employed shared title words and phrases, rather than shared MeSH terms, for bridging pairs of articles, for these and other reasons. It will be worth exploring how gaps identified using pairs of title terms (or pairs of terms appearing in abstract or full-text) will compare to those identified using pairs of MeSH terms.

Implementation

To provide a public test bed for studying gaps, we have implemented a tool called “Mine the Gap!” as part of the Anne O’Tate suite of value-added PubMed search tools (accessible with no login or passwords at http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi). The user carries out any topical PubMed query, after which a panel of options are displayed on the left hand of the page. One of these, the “Mine the Gap!” tool, identifies and characterizes the gaps with regard to a number of features (especially, article odds and author odds scores) and displays them in ranked form on the website. Optionally, the user can then click a button to automatically carry out Arrowsmith two-node searches on each gap, calculate the pR scores, and view the resulting output from each two-node search if desired. The dataset can be re-ranked on the website or exported as a comma-delimited text file for further study by informaticians, domain scientists or policy researchers.

The master sheet of gaps, features, and outcomes is attached to this article as Supplementary Table 2, and the list of gap-filling and non-gap-filling articles with their citations is attached as Supplementary Table 3.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

Supported by NIH P01AG039347. YP received a short-term fellowship from Nanjing University.

References

Adamic LA, Adar E. Friends and neighbors on the web. Social networks. 2003; 25(3):211–230.

Chen C, Chen Y, Horowitz M, Hou H, Liu Z, Pellegrino D. Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics. 2009; 3:191–209.

Chen, C. The Fitness of Information: Quantitative Assessments of Critical Evidence. New York: John Wiley & Sons; 2014.

Frank, E., Hall, MA., Witten, IH. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Fourth. Burlington, Massachusetts: Morgan Kaufmann; 2016. The WEKA Workbench.

Kastrin A, Rindflesch TC, Hristovski D. Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-based Discovery. Methods Inf Med. 2016; 55:340–346. DOI: 10.3414/ME15-01-0108 [PubMed: 27435341]

Larivière V, Haustein S, Börner K. Long-Distance Interdisciplinarity Leads to Higher Scientific Impact. PLoS ONE. 2015; 10:e0122565. http://doi.org/10.1371/journal.pone.0122565. [PubMed: 25822658]

Mishra S, Torvik VI. Quantifying Conceptual Novelty in the Biomedical Literature. D-Lib Magazine. 2016; 22:9–10. DOI: 10.1045/september2016-mishra

Packalen M, Bhattacharya J. Neophilia ranking of scientific journals. NBER Working Papers No 21579. 2015; doi: 10.3386/w21579

Radicchi F, Castellano C. Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. Journal of Informetrics. 2012; 6:121–130.

Smalheiser NR, Bonifield G. Two Similarity Metrics for Medical Subject Headings (MeSH): An Aid to Biomedical Text Mining and Author Name Disambiguation. J Biomed Discov Collab. 2016; 7:e1.doi: 10.5210/disco.v7i0.6654 [PubMed: 27213780]

Smalheiser NR, Torvik VI, Zhou W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. Comput Methods Programs Biomed. 2009; 94:190–197. DOI: 10.1016/j.cmpb.2008.12.006 [PubMed: 19185946]

Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab. 2008; 3:2.doi: 10.1186/1747-5333-3-2 [PubMed: 18279519]

Swanson DR. Undiscovered public knowledge. Library Quarterly. 1986; 56:103–118.

Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial intelligence. 1997; 91:183–203.

Theodosiou T, Vizirianakis IS, Angelis L, Tsaftaris A, Darzentas N. MeSHy: Mining unanticipated PubMed information using frequencies of occurrences and concurrences of MeSH terms. J Biomed Inform. 2011; 44:919–926. DOI: 10.1016/j.jbi.2011.05.009 [PubMed: 21684350]

Torvik VI, Smalheiser NR. A quantitative model for linking two disparate sets of articles in MEDLINE. Bioinformatics. 2007; 23:1658–1665. [PubMed: 17463015]

Uzzi B, Mukherjee S, Stringer M, Jones B. Atypical combinations and scientific impact. Science. 2013; 342:468–472. DOI: 10.1126/science.1240474 [PubMed: 24159044]

van Raan AFJ. Sleeping beauties in science. Scientometrics. 2004; 59:461–466.

Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. J Biomed Inform. 2009; 42:633–643. DOI: 10.1016/j.jbi.2008.12.001 [PubMed: 19124086]

Zhou J, Shui Y, Peng S, Li X, Mamitsuka H, Zhu S. MeSHSim: An R/Bioconductor package for measuring semantic similarity over MeSH headings and MEDLINE documents. Journal of bioinformatics and computational biology. 2015; 13:1542002. [PubMed: 26471719]

Table 1
Article sets obtained by querying PubMed

Shown are 10 disease-related and 10 biological processes queries. Each term was entered verbatim into PubMed (except for apoptosis[ti] where the term was restricted to the title field only, to keep the number of articles in the same range as the other queries). The first time slice is 1987-2005 and the second time slice is 2006-2010 inclusive. Unless otherwise noted throughout, t-tests were unpaired, 2-tailed, not assuming equal variance.

PubMed query	#articles 1st time slice	#articles 2nd time slice	fold increase
diseases			
acute myocardial infarction	33839	14567	1.430
alcoholism	31004	9412	1.304
AD	36807	18442	1.501
autism	7464	7478	2.002
colon cancer	46470	19990	1.430
cystic fibrosis	19395	7344	1.379
lupus	29735	11619	1.391
multiple sclerosis	23374	13438	1.575
schistosomiasis	7863	2412	1.307
suicide	28543	12059	1.422
mean	26449.4	11676.1	1.474
SD	12309.69	5310.90	0.203
biological processes			
alternative splicing	15616	6229	1.399
apoptosis[ti]	38518	21408	1.556
bacterial evolution	16140	15812	1.980
endocytosis	31775	16475	1.518
hyperpolarization	7534	2050	1.272
ion transport	31391	14749	1.470
meiosis	14067	5367	1.382
microtubules	16072	6765	1.421
protein aggregation	31932	13721	1.430
working memory	11792	9534	1.809
mean	21483.7	11211	1.524
SD	10745.55	6120.56	0.214
t-test	0.350	0.858	0.602

Table 2

Characteristics of the queries in terms of their gaps and MeSH terms

Significant differences by t-test are indicated in bold.

query	#articles 1st time slice	#mesh pairs	mesh pairs/article	#gaps	#gaps/article *100	#gaps filled	%gaps filled
diseases							
acute myocardial infarction	33839	326747	9.6559	51	0.1507	23	45.0980
alcoholism	31004	408042	13.1609	134	0.4322	29	21.6418
AD	36807	463576	12.5948	503	1.3666	158	31.4115
autism	7464	104235	13.9650	10	0.1340	5	50.0000
colon cancer	46470	767217	16.5099	758	1.6312	203	26.7810
cystic fibrosis	19395	364871	18.8126	282	1.4540	51	18.0851
lupus	29735	404202	13.5935	169	0.5684	55	32.5444
multiple sclerosis	23374	364889	15.6109	139	0.5947	43	30.9353
schistosomiasis	7863	132983	16.9125	56	0.7122	14	25.0000
suicide	28543	410629	14.3863	698	2.4454	59	8.4527
mean	26449.4	374739.1	14.5202	280	0.9489	64	28.9950
SD	12309.69	182364.79	2.58	275.27	0.75	64.77	12.20
biological processes							
alternative splicing	15616	468767	30.0184	21	0.1345	8	38.0952
apoptosis[ti]	38518	883333	22.9330	105	0.2726	42	40.0000
bacterial evolution	16140	363602	22.5280	161	0.9975	81	50.3106
endocytosis	31775	819690	25.7967	310	0.9756	104	33.5484
hyperpolarization	7534	214117	28.4201	142	1.8848	21	14.7887
ion transport	31391	887410	28.2696	242	0.7709	91	37.6033
meiosis	14067	295162	20.9826	345	2.4525	38	11.0145
microtubules	16072	388830	24.1930	85	0.5289	31	36.4706
protein aggregation	31932	851583	26.6686	889	2.7840	205	23.0596
working memory	11792	188437	15.9801	121	1.0261	47	38.8430
mean	21483.7	536093.1	24.5790	242.1	1.1827	66.8	32.3734
SD	10745.55	291073.45	4.19	248.84	0.90	57.68	12.26
t-test	0.3495	0.1580	**1.07E-05**	0.7505	0.5357	0.9198	0.5444

Table 3

Gap features in disease-related vs. biological processes article sets

See Materials and Methods for description of the features and outcomes. Significant differences by t-test are indicated in bold.

query	**Expected Query**	**Expected All**	**Co-occur All**	**Article Odds**	**Author Odds**
Diseases Mean	18.4652	931.5422	79.0424	0.1388	0.8911
SD	13.9621	1915.6879	236.2295	0.3230	1.2673
count	2800	2800	2783	2785	2785
Biol Proc Mean	16.3045	443.5984	82.8264	0.3041	1.1956
SD	9.2932	1100.8451	214.3451	0.4924	1.0071
count	2421	2421	2413	2413	2413
ttest	**3.06E-11**	**5.05E-30**	0.5451	**5.93E-44**	**8.03E-22**
query	**pR**	**CN score**	**AA score**	**MeSHSim**	
Diseases Mean	0.2520	278.8514	45.4896	0.0641	
SD	0.0819	134.6624	21.6283	0.1619	
count	1576	2800	2800	2800	
Biol Proc Mean	0.2544	436.9376	69.1662	0.1441	
SD	0.0864	170.1191	26.1958	0.2583	
count	1744	2421	2421	2421	
ttest	0.4159	**2.64E-260**	**4.86E-242**	**8.48E-39**	
query	**%filled**	**# of Articles**	**CPG**	**MCM**	
Diseases Mean	0.2286	0.3925	6.3745	0.7781	
SD	0.4200	0.9937	3.8246	1.1093	
count	2800	2800	640	640	
Biol Proc Mean	0.2759	0.5448	6.8721	0.9067	
SD	0.4471	1.5817	4.0109	1.7991	
count	2421	2421	668	668	
ttest	**8.81E-05**	**4.37E-05**	**0.0218**	0.1180	

Table 4

Gaps across disciplines in biological processes article sets showing some of their relevant features.

query	Gaps	Article Odds	Author Odds	pR
apoptosis	Liver Neoplasms::Neurons	0.0065	0.4047	0.499
apoptosis	Lung Neoplasms::Neurons	0.0524	0.3859	0.462
apoptosis	Colonic Neoplasms::Neurons	0.0214	0.4958	0.451
apoptosis	Leukemia::Neurons	0.0175	0.3992	0.424
apoptosis	Leukemia::Rats, Sprague-Dawley	0.0276	0.4482	0.415
bacterial_evolution	Bacterial Typing Techniques::Mitochondria	0.0143	0.4193	0.414
meiosis	Infertility, Male::Saccharomyces cerevisiae	0.0465	0.3794	0.411
meiosis	Genes, Plant::Spermatogenesis	0.0529	0.4472	0.411
meiosis	Genes, Fungal::Infertility, Male	0	0.3626	0.403
working_memory	Magnetic Resonance Imaging::Rats, Inbred Strains	0.077	0.377	0.398
meiosis	Genes, Plant::Testis	0	0.273	0.388
meiosis	Schizosaccharomyces::Swine	0.1713	0.4477	0.388
ion_transport	Cystic Fibrosis::Rats, Wistar	0.0389	0.3808	0.386
meiosis	Chromosome Aberrations::Gene Expression Regulation, Fungal	0.1012	0.4564	0.375
working_memory	Rats, Inbred Strains::Verbal Learning	0.0091	0.2497	0.375
meiosis	Follicle Stimulating Hormone::Saccharomyces cerevisiae	0.0195	0.264	0.372
meiosis	Genes, Plant::Spermatozoa	0	0.2596	0.366
working_memory	Macaca mulatta::Reading	0	0.4729	0.362
working_memory	Rats, Wistar::Verbal Learning	0.031	0.2003	0.357
meiosis	Genes, Fungal::Swine	0.083	0.4871	0.356
working_memory	Rats, Sprague-Dawley::Verbal Learning	0	0.2442	0.355
protein_aggregation	Coronary Disease::Rats, Wistar	0.0482	0.4354	0.351

Table 5

Multiple linear regression to predict the number of gap-filling articles appearing in the second time slice in the combined dataset

Shown are feature weights and performance values for 10-fold cross-validation of the entire set of 5221 gaps over 20 queries. Note that features having skewed distributions were converted to square root values to improve quasi-normality. Also note that the CN feature was removed as being redundant with AA (see text).

Weights of each feature:	
0.0338 * sq exp query +	
−0.0011 * sq exp all +	
0.0125 * sq cooccur all +	
0.4686 * sq article odds +	
−0.1208 * sq author odds +	
0.7056 * pR +	
0.0045 * AA score +	
−0.1124 * MeSHSim +	
−0.3392	
Correlation coefficient	0.4174
Kendall’s tau	0.3207
Spearman’s rho	0.4001
Mean absolute error	0.4066
Root mean squared error	0.5463
Relative absolute error	85.0147 %
Root relative squared error	90.8695 %

Table 6

Multiple linear regression to predict the number of gap-filling articles appearing in the second time slice in the 10 disease queries only

As in Table 5 but only for the 2800 gaps in the 10 disease related queries.

Weights of each feature:	
0.0301 * sqrt exp query +	
−0.0034 * sqrt exp all +	
0.0222 * sqrt cooccur all +	
0.3411 * sqrt article odds +	
1.0526 * pR +	
0.0048 * AA score +	
−0.2386 * MeSHSim +	
−0.4493	
Correlation coefficient	0.4712
Kendall’s tau	0.3593
Spearman’s rho	0.4465
Mean absolute error	0.3636
Root mean squared error	0.4923
Relative absolute error	82.8104 %
Root relative squared error	88.1807 %

Table 7

Multiple linear regression to predict the number of gap-filling articles appearing in the second time slice in the 10 biological processes queries only

As in Table 5 but only for the 2421 gaps in the 10 biological processes queries.

Weights of each feature:	
−0.0058 * Expected Query +	
0.0771 * sqrt exp query +	
0.0001 * Expected All +	
−0.004 * sqrt exp all +	
0.7522 * sqrt article odds +	
−0.366 * sqrt author odds +	
0.53 * pR +	
0.0009 * AA score +	
−0.3088	
Correlation coefficient	0.4059
Kendall’s tau	0.3036
Spearman’s rho	0.3805
Mean absolute error	0.4491
Root mean squared error	0.5894
Relative absolute error	86.3533 %
Root relative squared error	91.3554 %