Supplementary Information: Examining cohesion and diversity in collaboration networks of pharmaceutical clinical trials

Gary Lin, Sauleh Siddiqui, Jen Bernstein, Diego Martinez, Lauren Gardner, Tenley Albright, & Takeru Igusa

July 11, 2020

S1 Data and Processing

The analysis was performed on the Aggregate Analysis of ClinicalTrial.gov (AACT) database from the Clinical Trials Transformation Initiative (CTTI) at Duke University Tasneem et al. (2012) and the BioMedTracker Pharma Intelligence Database from Informa LLC Thomas et al. (2016). The AACT database was accessed on January 12, 2017, and the Biomed Tracker database was accessed on January 15, 2017.

We used the intersection of the two databases where data existed for the variables of interest. Each distinct trial is defined as a unique national clinical trial (NCT) identifier number. The AACT database provided information on collaboration based on *lead sponsor* and *collaborators*. Hence, we were able to connect each unique NCT number with a set of organizations that were affiliated with that particular trial to construct our collaboration network. The AACT also provided the official start and end dates of each trial. We selected the trials that were designated as a “drug” intervention defined by ClinicalTrials.gov. Our analysis is refined to 4,494 organizations[1] and 18,040 trials.

[1] In our analysis, we consider each unique name to be a distinct actor. Therefore, subsidiaries that have a different name from the parent company were considered to be separate organizations.

From the BioMedTracker database, we were able to extract the therapeutic research area through *disease group* designation and FDA approval status of each trial. In our analysis, we considered 21 different therapeutic areas. The disease group data were used to calculate knowledge mix, research diversification, and collaboration diversity indices, which are defined along with other regression variables in Section S3.

Additional data collection was conducted using a combination of manual web searches and text-mining to classify each actor into six categories: Academic, Government, Nonprofit, Industry, Hospital System, or Large Pharmaceutical. Section S3.2 outlines our method of classification. Actor classification was done to aid our regression analysis since the roles of different types of organization vary within a collaboration network. For example, academic institutions usually collaborate differently than nonprofit organizations.

S2 Constructing the Collaboration Network

Using the *collaborator* section of the AACT database, we structured the data so each clinical trial, which is defined as a unique NCT number is connected to a set of collaborators and a lead sponsor. From this structured data, we constructed an undirected two-mode affiliation network that connected lead sponsors and collaborators to a corresponding clinical trial based on their involvement at that particular time. The network is called two-mode because there are two types of nodes: actors and clinical trials. Using a bipartite projection, the two-mode network is converted into an undirected, one-mode collaboration network with only actor nodes to better capture the structural features and relationship between each actor. The one-mode network is represented as a set of actors that are actively conducting clinical trials at that particular time. A link exists between a pair of actors that collaborate on at least one clinical trial. The network and regression analysis were performed on the one-mode network. Figure S1 shows the translation between a two-mode and one-mode network.

We considered the network dynamics by building a distinct network for each month which essentially is a "snapshot" of all active clinical trials and their associated collaborations. A clinical trial is considered active if the observed time period is after the *start date* and before the *end date* as listed in the AACT database. We utilized the Biomed Tracker database to determine whether a trial was suspended (failed) or is associated with a treatment that was approved by the U.S. Food and Drug Administration (success). A node only exists if the actor is involved in at least one clinical trial at the observed time period. Hence, the number of nodes differs between time periods.

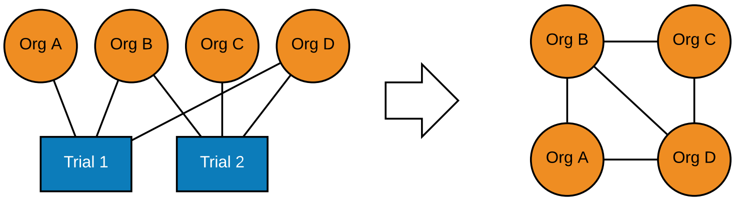


Figure S1: Bipartite projection of two-mode affiliation network to a weighted one-mode collaboration network.

S3 Definitions of Regression Variables

We discuss our methodology in calculating the regression variables in this section. The regression variables are divided into five categories: (1) organizational type, (2) expertise, (3) structural measures, (4) organizational measures, and (5) collaboration measures. The summary statistics of variables that were considered structural, organizational, and collaboration measures that were used in the regression analysis are in Table S5 of Section S5.

Table S1: List of Variables

Variables	Definition	Type
δ_{it}	Number of degrees of actor i at time t	Structural Measure
σ_{jk}	Number of shortest path between nodes j and k	Structural Measure
$\sigma_{jk}(i)$	Number of shortest paths from node j to node k that goes through node i	Structural Measure
BT_{it}	Betweenness centrality of actor i at time t	Structural Measure
CC_{it}	Local clustering coefficient of actor i at time t	Structural Measure
CD_{it}	Collaboration diversity of actor i at time t	Collaboration Measure
KD_{ijt}	Knowledge distance between actors i and j at time t	Collaboration Measure
$\langle KD\rangle_{it}$	Average knowledge distance of actor i at time t	Collaboration Measure
L_{it}	Number of links of actor i at time t	Structural Measure
n_{idt}	Number of clinical trials conducted by actor i in therapeutic area d at time t	Organizational Measure
p_{idt}	Number of collaborators of actor i with expertise in disease groupt d at time t	Collaboration Measure
RD_{it}	Research diversification of actor i at time t	Organizational Measure
$\langle RD\rangle_{it}$	Average neighbor research diversification of actor i at time t	Collaboration Measure
x_{idt}	Fraction of clinical trials conducted by actor i in therapeutic area d at time t	Organizational Measure
$\mathbf{x}_{it}$	Knowledge mix of actor i at time t	Organizational Measure
z_{idt}	Fraction of collaborators of actor i with expertise in disease group d at time t	Collaboration Measure
d	Disease	Index
i, j, k	Actor (node)	Index
t	Time	Index
D	Disease group set	Set
$E(i,j)$	Link set between actors i and j	Set
V	Node set	Set

S3.1 Organizational Types

Each actor is classified into one of six *organization types*: Academic, Government, Nonprofit, Industry, Hospital System, or Large Pharmaceutical. The organization types of each actor are defined in Table S2. The organization type designation for each actor is consistent for all time periods. For each of the 4,494 actors, we were able to classify the actors based on publicly available information on the Web regarding their organizational function and mission. From that information, we were able to categorize many of the actors based on their registered names in the AACT database using expert judgment. In Table S2, we show the inclusion criteria for determining the organization types. We differentiated between large pharmaceutical companies and industry actors by selecting the companies that were ranked as either top 25 with the highest market capitalization in 2016, top 15 revenue in 2016, or top 15 R&D budgets in 2016 (see Table S2). The complete listing for large pharmaceutical companies is included in Table S3.

Table S2: Classifications of Organization Types

Category	Description and Inclusions Criteria	Count
Academic	University, colleges, medical schools, academic health centers, faculty investigators, and learning hospitals.	639
Government	National and federal institutes, federal and state agencies, national ministries, regulators, and veteran hospitals	123
Nonprofit	Patient advocacy, trusts, initiatives, non-federal research institutes, and collaborative groups	398
Industry	Pharmaceutical/biotechnology firms (that are not considered large pharmaceutical companies), corporations, multinationals, holdings, and private clinics	2989
Hospital System	Healthcare networks, non-learning hospital, and community hospitals	255
Large Pharmaceutical Companies	Top 25 companies with highest market capitalization, top 15 revenue, or top 15 R&D Budgets from 2016 (See Table S3 for complete listing large pharmaceutical companies)	90

S3.2 Expertise

Each trial in the BioMedTracker database is associated with a therapeutic area based on the main treatment objectives. In our sample from the BioMedTracker database, there are 21 different therapeutic areas. Figure S2 shows the distribution of trials with respect to the therapeutic designation.

We assumed that the experience and data gained from a clinical trial contributes to the knowledge regarding the therapeutic area. For each actor, we were able to designate them as an "expert" in a therapeutic area if most of their clinical trial participation was in that particular therapeutic area. Although all actors have one expertise, they can be still be diversified with less concentration in their expert therapeutic area. This is described by an actor's research diversification (see Section S3.4.2).

Table S3: Actors that are considered Large Pharmaceutical Companies

Stryker Orthopaedics
Johnson & Johnson
Amgen Research Munich GmbH
Alexion Pharma GmbH
Abbott Medical Optics
Teva Neuroscience, Inc.
Johnson & Johnson Medical, China
Astellas Pharma Europe B.V.
Abbott Products
Shire Regenerative Medicine, Inc.
Merck Serono Co., Ltd., Japan
Sanofi-Synthelabo
Teva Women's Health
Merck Serono S.A., Geneva
Johnson & Johnson Pte Ltd
Teva Pharmaceuticals USA
MAP Pharmaceuticals, Inc., a wholly-owned subsidiary of Allergan
Stryker Biotech
Stryker Instruments
Daiichi Sankyo UK Ltd.
Sanofi Pasteur MSD
Stryker Nordic
Durata Therapeutics Inc., an affiliate of Allergan plc
Abbott Diabetes Care
Roche-Genentech
TEVA
Teva Pharma
Abbott Japan Co., Ltd
Astellas Pharma Global Development, Inc.
Allergan Medical
Abbott Diagnostics Division
Astellas Pharma Europe Ltd.
Celgene
Stryker MAKO Surgical Corp
Merck Serono S.A., Switzerland
Astellas Pharma China, Inc.
Stryker MAKO Corp
Genentech/Roche
Johnson & Johnson Medical Companies
Bristol Meyers Squibb BMS
Teva Pharmaceutical Industries, Ltd.
Janssen/GlaxoSmithKline
Janssen, GlaxoSmithKline GSK
Astellas Pharma Korea, Inc.
Stryker South Pacific

AstraZeneca
Eli Lilly and Company
Johnson & Johnson Pharmaceutical Research & Development, L.L.C.
Sanofi
Wyeth is now a wholly-owned subsidiary of Pfizer
Hoffmann-La Roche
Biogen
Bristol-Myers Squibb
Amgen
Pfizer
Novartis Pharmaceuticals
GlaxoSmithKline
Celgene Corporation
Astellas Pharma Inc
Astellas Pharma US, Inc.
Bayer
Genzyme, a Sanofi Company
Abbott
Gilead Sciences
Shire
Novartis
Roche Pharma AG
Novo Nordisk A/S
Alexion Pharmaceuticals
Merck KGaA
Daiichi Sankyo Co., Ltd.
Takeda
Daiichi Sankyo Inc.
Allergan
Abbott Vascular
Teva Pharmaceutical Industries
Vertex Pharmaceuticals Incorporated
Orthovita d/b/a Stryker
AbbVie prior sponsor, Abbott
Stryker Neurovascular
Regeneron Pharmaceuticals
Boehringer Ingelheim
Teva Branded Pharmaceutical Products, R&D Inc.
Sanofi Pasteur, a Sanofi Company
AbbVie
Shire Human Genetic Therapies, Inc.
Janssen, LP
King Pharmaceuticals is now a wholly-owned subsidiary of Pfizer
Gamida Cell -Teva Joint Venture Ltd.
Novartis Vaccines

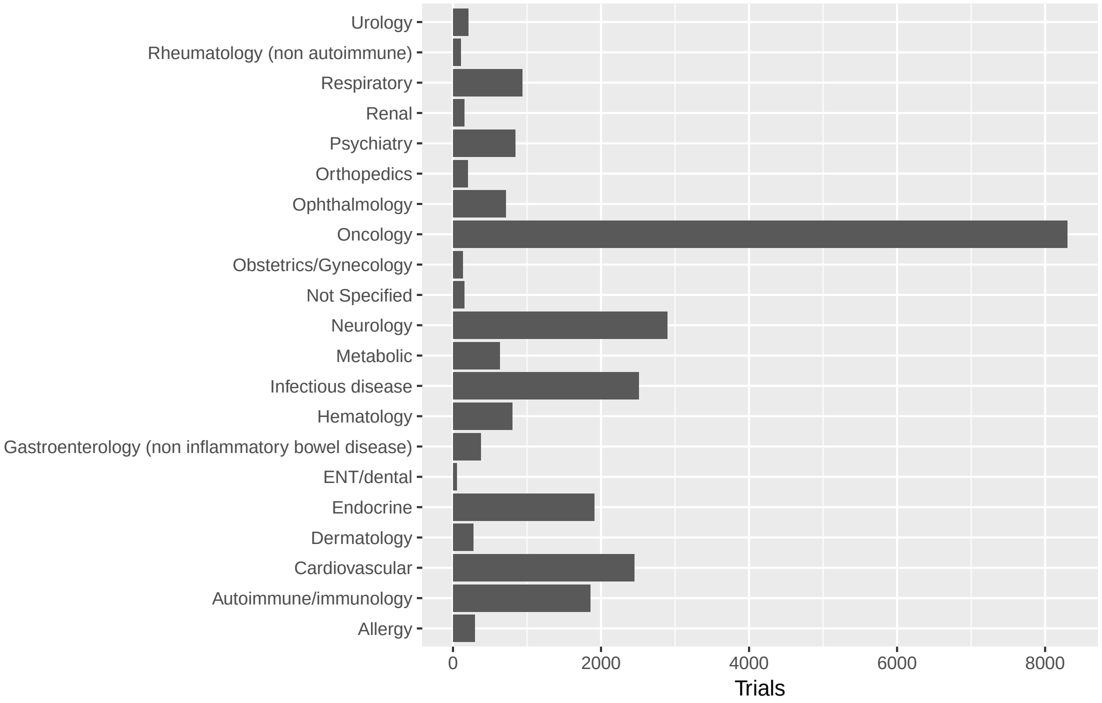


Figure S2: The distribution of trials in our analysis is shown by therapeutic disease groups from Jan 2006 - Jan 2016

S3.3 Structural Measures

We define structural measures as metrics related to graphical properties of the collaboration network. These measures include the betweenness centrality and clustering coefficient. These variables were calculated for each time step based on our dynamic collaboration network and were included in our regression analysis for each actor.

S3.3.1 Betweenness Centrality

The *betweenness centrality* BT_{it} for actor i at time t is defined as

$$BT_{it} = \sum_{j,k \in V} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \quad s.t.\ j \neq k \tag{1}$$

where the denominator σ_{jk} represents the shortest path between nodes j and k in node set, V, which includes all possible pariwise combination of nodes in the network, and the numerator $\sigma_{jk}(i)$ represents the number of shortest paths from node j to node k that goes through node i Freeman (1980). This metric is useful for measuring the extent to which a node acts as a "bridge" between two communities.

The motivation for including betweenness centrality in our analysis is based on previous literature Granovetter (1973) that actors that have a high betweenness centrality in the clinical trials collaboration network are in positions that would enable them to be a conduit for knowledge flow. Furthermore, these organizations that are in central positions with a high betweenness centrality are also able to tap into the different knowledge bases of a variety of actors. In our analysis, large pharmaceutical organizations generally have a high betweenness centrality. Actors that have low betweenness centrality are typically on the peripheries of a network, such as biotechnology and life science startups. The betweenness centrality was calculated using the *igraph* package Csardi and Nepusz (2006) in R R Core Team (2018) that utilizes the algorithm developed by Brandes Brandes (2001).

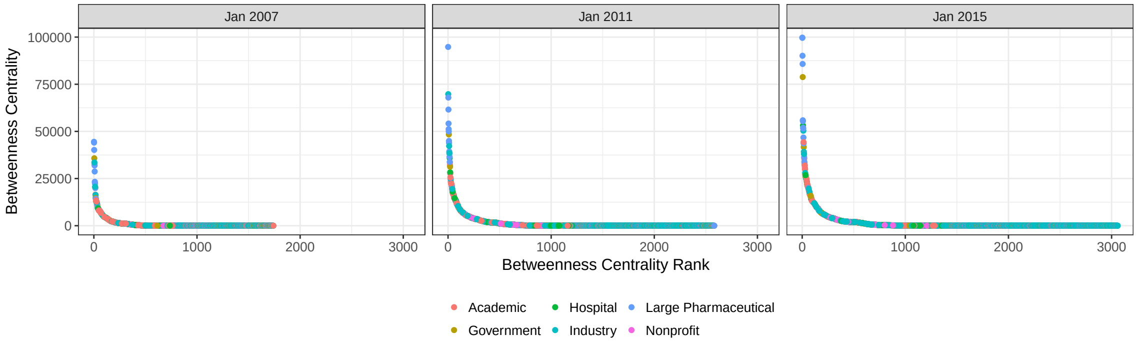


Figure S3: Rank-size plot of betweenness centrality for Januarys of 2007, 2011, and 2015.

S3.3.2 Local Clustering Coefficient

The *local clustering coefficient* CC_{it} measures the extent to which an actor's neighbors are connected to each other for actor i at time t. If an actor's local neighborhood, which includes itself and its neighbor, is fully connected as a clique (fully-connected subgraph), then the clustering coefficient would be one, while a completely unconnected local network would be zero. Formally, the local clustering coefficient for an undirected graph is defined as

$$CC_{it} = \frac{2L_{it}}{\delta_{it}(\delta_{it} - 1)} \tag{2}$$

where L_{it} represents the number of links between the neighbors of actor i at time t, and δ_{it} represents the number of degrees of actor i at time t.

The local clustering coefficient is a good indicator of local cohesiveness Guler and Nerkar (2012). The local clustering coefficient describes the level to which a local neighborhood of an actor is a clique. In the context of collaboration trials, actors that have high measures of local clustering tend to speed up knowledge transfer within their local network Schilling and Phelps (2007). However, local clustering can also result in knowledge redundancy in which organizations are essentially stuck in an echo chamber of redundant knowledge. All actors' local clustering coefficient was calculated using the *igraph* package Csardi and Nepusz (2006) in R R Core Team (2018).

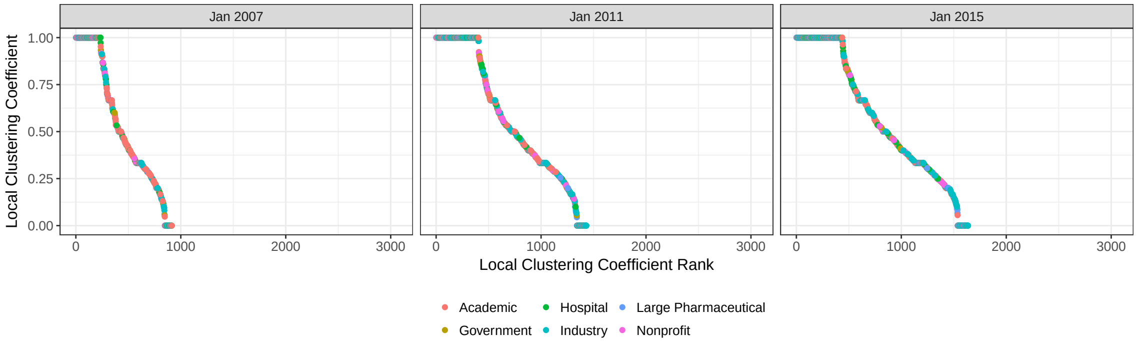


Figure S4: Rank-size plot of local clustering coefficient for Januarys of 2007, 2011, and 2015. Nodes with less than 2 neighbors are removed.

S3.3.3 Vertex Degree

Per the conventional definition of *degree centrality*, this is simply the number of links that are adjacent to the observed node, which we will define as δ_{it}. Also known as vertex degree, δ_{it} is the degree count at time t for actor i. Please note that we did not include this measure in the regression analysis because of its high collinearity with cumulative trials conducted.

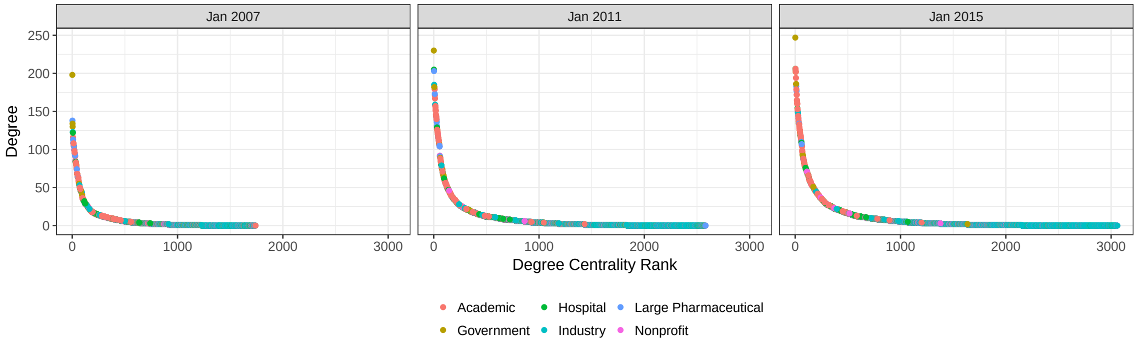


Figure S5: Rank-size plot for degree centrality for Januarys of 2007, 2011, and 2015. Nodes with less than 2 neighbors are removed.

S3.4 Organizational Measures

Organizational measures include variables that relate to the actors' research experience. The variables that we introduce include knowledge mix and research diversification.

S3.4.1 Knowledge Mix

Similar to the knowledge space index was developed by Vaccario et al. (2018) to show the distribution of research in different industrial patent areas, our *knowledge mix* index shows the distribution of an actor's research portfolio in each therapeutic research area. They considered knowledge to be demonstrated when a company files a patent in a certain industrial area (e.g. aerospace, pharmaceutical, manufacturing), while we considered the knowledge mix to be trials that were conducted in each disease group area. We utilize the knowledge mix index to determine the amount of experience gained when an organization conducts research in a particular disease domain. As shown in the main article, the knowledge mix for actor i at time t is defined as a vector $\mathbf{x}_{it}$ with the element x_{idt} where d is the therapeutic research area. The element x_{idt} is defined as

$$x_{idt} = \frac{n_{idt}}{\sum_{d \in D} n_{idt}} \tag{3}$$

If actor i does not have any approvals at time t, then the knowledge mix is a null vector $\mathbf{x}_{it} = \mathbf{0}$. We define experience (i.e. knowledge) as n_{idt} which represents the number of clinical trials in the therapeutic area d that actor i has been involved as a sponsor or partner at time t.

Since the knowledge mix is not a scalar, we did not use the variable directly in our regression analysis. Instead, the knowledge mix vector is used to calculate research diversification, mean neighbor research diversification, knowledge distance, and collaboration diversity.

S3.4.2 Research Diversification

On a firm level, an organization may decide to adopt two approaches: broadly diversify in different therapeutic disease areas (jack-of-all-trades) or specialize in one therapeutic research area (master-of-one). The decision to diversify is usually driven by the size of organizations that determines the economies of scope Cockburn and Henderson (2001). Larger companies tend to be more diverse than smaller companies since they have the resources to absorb smaller companies Makri et al. (2010). There are certain trade-offs and advantages for adopting each strategy which may vary depending on organizational goals.

For each actor, we quantified *research diversification* using an entropic measure that measures the heterogeneity of actor i's knowledge portfolio. This is also known as a technological distance in economic literature Bar and Leiponen (2012).

$$RD_{it} = \sum_{d \in D} x_{idt} \ln \left(\frac{1}{x_{idt}} \right) \tag{4}$$

where x_{idt} is the element of the knowledge mix vector that represents the percentage of clinical trials experience in disease d at time t. This measure gives us an impression of the level of interdisciplinary in an organization's research portfolio. We assume that a company that has completed no trials will have a research diversification of zero, which would be equal to companies that have only one trial.

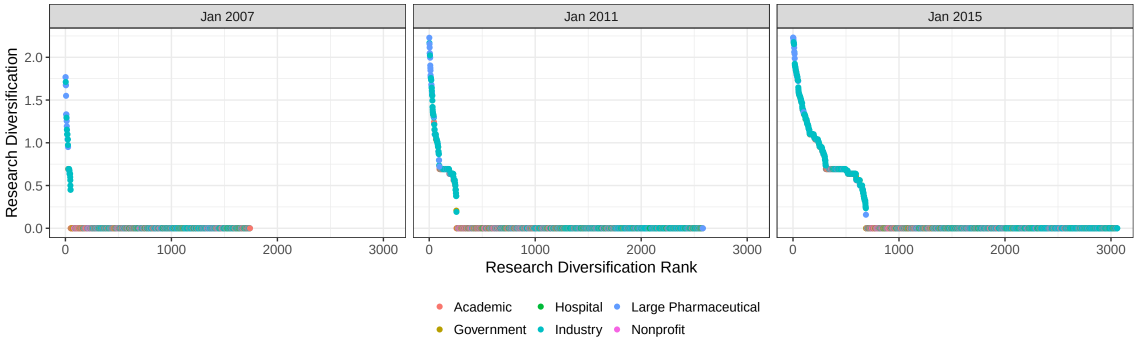


Figure S6: Rank-size plot of research diversification for Januarys of 2007, 2011, and 2015.

S3.5 Collaboration Measures

In this section, we introduce mean knowledge distance, collaboration diversity, and mean neighbor research diversity. These variables relate to an actor's collaboration with partnering actors.

S3.5.1 Mean Knowledge Distance

Based on Vaccario et al. (2018)'s definition of *knowledge distance* as the Euclidean distance between organizations i and j at time t. In other economic literature, this is known as the technological distance Bar and Leiponen (2012). This is formally defined as

$$KD_{ijt} = \|\mathbf{x}_{it} - \mathbf{x}_{jt}\| = \sqrt{\sum_{d \in D} (x_{idt} - x_{jdt})^2}, \tag{5}$$

where x_{idt} represents an element of the knowledge mix vector $\mathbf{x}_{it}$ that was defined in S3.4.

This link-specific metric was meant to compare the differences between the two firms' patent portfolios. However, we have adopted this metric to measure the differences between pairs of the organization's research portfolio which is represented by the knowledge mix vector. The knowledge distance is at a maximum ($KD = \sqrt{2}$) when actors are concentrated in two different therapeutic areas. When two firms are concentrated in the same therapeutic area, the knowledge distance equals to 0 because they are identical in expertise.

One of the properties of the Euclidean-based knowledge distance in (5) is that the measure takes into account the research diversification of an actor. Let's say actor 1's knowledge mix is solely concentrated in Neurology, actor 2's knowledge mix is solely concentrated in Urology, and actor 3's knowledge mix is divided between Oncology and Urology. In this situation, the knowledge distance between actor 1 and actor 2 is larger than actor 1 and actor 3 even though both actors 2 and 3 are in exclusive research fields relative to actor 1. This is a well-known property of Euclidean Distances and fits our case since we are implying that more interdisciplinary actors have more capacity to function in other fields than specialists.

In our analysis, we calculated the *mean knowledge distance* $\langle KD \rangle_{it}$ for all incident links to actor i at time t. This allows us to aggregate all the knowledge distances (links) for each actor into one actor-specific scaler variable that can be used in the regression. We can define this as

$$\langle KD \rangle_{it} = \frac{\sum_{j \in E(i,j)} KD_{ijt}}{\delta_{it}} \quad s.t.\ i \neq j \tag{6}$$

where δ_{it} is the number of degrees for actor i and E represents the set of links connected to actor i at time t.

S3.5.2 Collaboration Diversity

We created a metric called *collaboration diversity* to measure the level of interdisciplinary collaboration that an actor is engaged in. Given that all actors in our network are classified as an expert

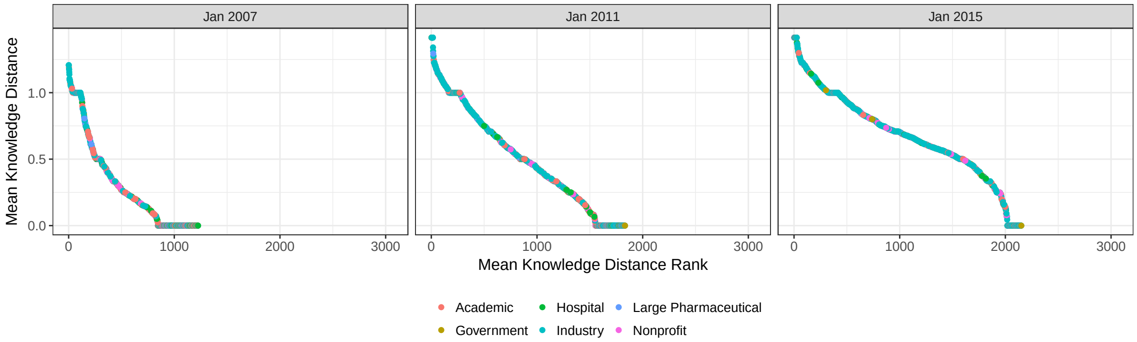


Figure S7: Rank-size plot of mean knowledge distance for Januarys of 2007, 2011, and 2015.

in a single therapeutic area, we can quantify the heterogeneity of an actor's set of collaborators in terms of expertise using a similar entropic measure like research diversification that we defined in Section S3.4. The reader should note that collaboration diversity does not consider the neighboring actor's research diversity, only their expertise. We consider the neighbor's research diversification measure in Section S3.5.3.

By simply identifying the elements of the knowledge mix vector that has the largest value, we can determine the therapeutic area that an actor would be considered an "expert."[2] Based on this method, we were able to designate each actor in our graph with one therapeutic expertise and calculate the level of diversity of collaborators for each actor. This measure is similar to the entropic measure in (4).

For each actor i, there exist a set of collaborators (neighbors) that each have an expertise. We define z_{idt} as being equal to the fraction of collaborators that are experts in therapeutic area d. We can express this mathematically,

$$z_{idt} = \frac{p_{idt}}{\sum_{d \in D} p_{idt}} \tag{7}$$

where p_{idt} is the number of experts in disease group D that actor i collaborate with on a clinical trial at time t. Since z_{idt} is a fraction, we can formally define our collaboration diversity as an entropic measure in Equation (8). For actors that have no collaborators, we assume $CD = 0$.

[2] For some of the actors, they have expertise in 2 or more therapeutic areas because there are multiple elements in the knowledge mix vector that are equal and have the largest values (e.g. $\max_{d \in D} x_{id} = x_{i1} = x_{i2}$). Around 5% of actors for the entire dataset fall under this classification. In these cases, we designate the actors as a separate group from the rest of the therapeutic designations. Partners with no expertise in any therapeutic areas are excluded.

$$CD_{it} = \sum_{d \in D} z_{idt} \ln\left(\frac{1}{z_{idt}}\right) \tag{8}$$

This metric examines the set of partners for each actor and measures the diversity of expertise. In essence, collaboration diversity characterizes the breadth of expertise in the actor's collaboration network. Actor A's collaboration diversity would be ~ 0.64 if they worked with 2 experts in neurology and 1 expert in oncology. Actor B would have a collaboration diversity of ~ 3 if they conducted 20 trials with 20 different experts all specializing in different therapeutic areas. In this case, we would say Actor B is more diverse in their collaboration.

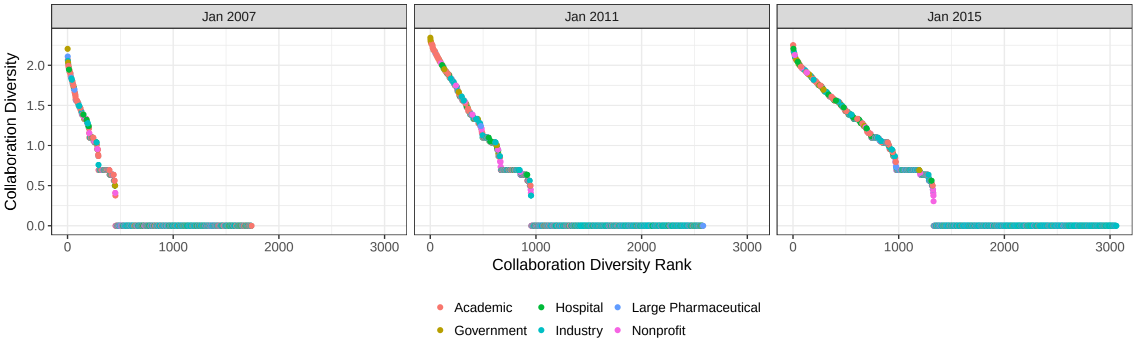


Figure S8: Rank-size plot of collaboration diversity for Januarys of 2007, 2011, and 2015.

S3.5.3 Mean Neighbor Research Diversification

From the research diversification that is defined in (4), we can determine *mean research diversification* of all the neighboring organization of actor i at any given time period t. The mean research diversification $\langle RD \rangle_{it}$ is simply

$$\langle RD \rangle_{it} = \frac{\sum_{j \in E(i,j)} RD_{jt}}{\delta_{it}}. \tag{9}$$

where δ_i is the number of degrees for node i.

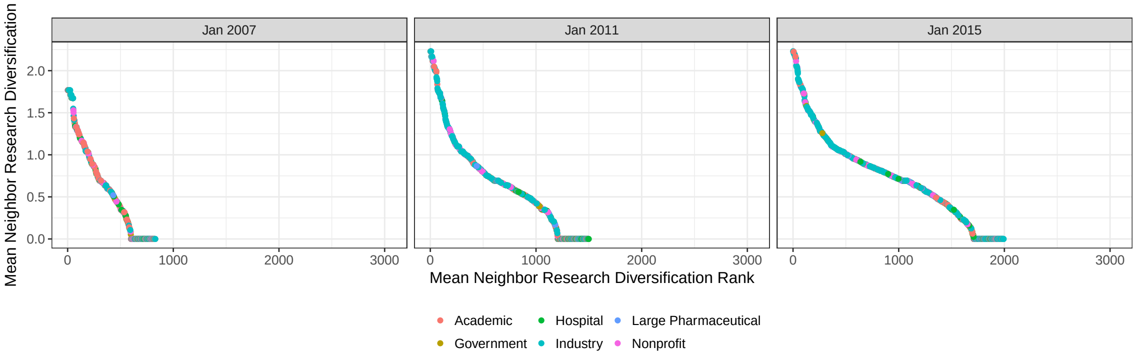


Figure S9: Rank-size plot for mean neighbor research diversification for Januarys of 2007, 2011, and 2015.

S4 Multivariate Regression Analysis

Using the set of 121 temporal collaboration networks that were constructed, we created a panel dataset composed of the structural measures of the collaboration network (Section S3.3), organizational measures (Section S3.4), and collaboration measures (Section S3.5). We plotted the trends for all of the months in Figure S12. For our regression analysis, we discretized the data based on 6-month intervals, starting in January 2006 and ending in January 2016 (19 time periods).

In order to obtain explanatory relationships of each network measure on an actor's ability to develop successful treatments, we utilized a negative-binomial generalized linear model (GLM) and a beta regression model to respectively predict *cumulative trial successes* that represents research output, and *trial success rate* that represents research efficiency. Given that the number of cumulative trial successes is a count variable and its variance is overdispersed, we selected the negative-binomial regression over a Poisson regression. For the variable, trial success rate, a beta regression was selected to predict a response variable with a continuous unit interval (0,1).

We also considered 3 different lag lengths: 1, 2, and 5 years to analyze delayed effects. In the main article, we presented a set of **comprehensive models** that includes all relevant measured variables. In addition to the comprehensive model, we developed two sets of models: reduced and base models. **Reduced models** include heuristically selected explanatory variables from the comprehensive model in order to reduce collinearity. Specifically, we eliminated variables with high variational inflation factors (VIF) and low P-values. **Base models** only include control variables that were not related to any of our developed metrics in Section S3 – this includes previous experience, previous success, and cumulative trials conducted.

The results for the base model (control variables), reduced model (selected variables), and comprehensive model (all variables) are shown in Tables S6 - S8 and S10 - S12. For each predicted variable and lag length, we show the set of three models side-by-side to demonstrate the robustness of our regressions.

For all models, we accounted for confounding temporal factors (e.g. regulations, policy changes) by adding fixed effects for each time period. We also included fixed effects for each organizational class (e.g. academic, nonprofit) to capture the differences in actors' behaviors.

The negative-binomial and beta regressions will have the following generalized form,

$$g(y_{it}) = \beta_0 + \beta_1 x_{1i(t-k)} + \beta_2 x_{2i(t-k)} + \ldots + \beta_{Mi(t-k)} + \gamma_t + \kappa_i + \epsilon_{it} \tag{10}$$

where $g(\cdot)$ is the link function, γ_t is the fixed effect of each 6-month time interval between 2006-2016, κ_i is the fixed effect for organizational type, and ϵ_{it} is the error term. There are M covariates (attributes) designated as x_{jit} that measures attribute j of actor i at time t. We employed the statistical software R R Core Team (2018) to estimate the coefficients β using the method of maximum likelihood to regress against the response variable y_{it}.

S4.1 Cumulative Trial Successes Regression

For the comprehensive, reduced, and base models, we utilized a negative-binomial regression class of GLMs to estimate the linear relationship of cumulative trial successes CTS_{it} for all actors i at time t. The variable, cumulative trial successes, is meant to capture the research output of an actor.

We standardized all count and continuous variables such that the mean is zero and the standard deviation is one. This rescaling allows us to ignore units and have a better comparison of regression coefficient values. The standardized variables include cumulative trials conducted $Trials_{it}$, collaboration diversity CD_{it}, mean knowledge distance $\langle KD \rangle_{it}$, local clustering coefficient CC_{it}, betweenness centrality BT_{it}, research diversity RD_{it}, and mean neighbor's research diversity $\langle RD \rangle_{it}$ in the comprehensive model.

For our main analysis, our comprehensive model is defined as

Comprehensive Model (All Variables):

$$\begin{aligned} \log(CTS_{it}) = \beta_0 &+ \beta_1 PrevSucc_{i(t-k)} + \beta_2 Trials_{i(t-k)} + \beta_3 CD_{i(t-k)} + \beta_4 \langle KD \rangle_{i(t-k)} \\ &+ \beta_5 CC_{i(t-k)} + \beta_6 BT_{i(t-k)} + \beta_7 RD_{i(t-k)} + \beta_8 \langle RD \rangle_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it} \end{aligned} \tag{11}$$

To account for the delayed effect, the dependent variable is lagged, and all the independent variables are taken from an earlier period in our panel dataset with lag differences $k = 1$, 2, and 5 years. The dummy variables γ_t and κ_i characterize the fixed effects of time and organizational class as defined in Table S2. For more information on these variables, refer to Section S3. The coefficient estimates and standard errors for (11) are listed as Models A1-3, A2-3, and A3-3 in Tables S6 - S8.

In our analysis, we also accounted for reverse causality in which the network structure is the result of the organizational success of the company which makes the network metrics endogenous. For instance, organizations may want to work with a more successful company (i.e., preferential attachment), resulting in more trials. We attempted to control the endogenous effects by including past success as a dummy variable $PrevSucc_{i(t-k)}$ to account for confounding factors between network structure and cumulative success in trials. In our model, $PrevSucc_{t-k} = 1$ when a company has participated in at least one successful clinical trial at or before time $t-k$, otherwise $PrevSucc = 0$.

To illustrate the robustness of our model to explain cumulative trial successes, two other set of regression models were developed in addition to (11) to explain the relative effects of all the variables: base model and reduced model. The base model only includes the control variables, and the results are listed under Models A1-1, A2-1, and A3-1 in Tables S6 - S8. (12) shows the base model's mathematical structure.

Base Model (Control Variables):

$$\log(CTS_{it}) = \beta_0 + \beta_1 PrevSucc_{i(t-k)} + \beta_2 Trials_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it} \tag{12}$$

The reduced models that include all the variables, and the results are listed as Models A1-2, A2-2, and A3-2 in Tables S6 - S8. The reduced model was developed based on variable selection methods from the comprehensive model presented in (11). Variable selection was done using the following heuristic. The purpose of this heuristic is to eliminate variables that might be collinear. For example, the heuristic removes betweenness centrality, BT which is collinear to the number of cumulative trials, $Trials$, with a correlation of 0.62 for one-year lag.

1. Based on the comprehensive model with a 1-year lag, we eliminated all variables with P-values more than 0.001.

2. Using the method of random holdouts with a 30-70% test-to-training set ratio, we manually eliminated variables until the predictive accuracy was the highest based on mean squared error (MSE) between the trained model and test set. We did not eliminate any variables associated with the fixed effects.

The resulting reduced model from our variable selection is presented in (13).

Reduced Model (Selected Variables):

$$\log(CTS_{it}) = \beta_0 + \beta_1 PrevSucc_{i(t-k)} + \beta_2 Trials_{i(t-k)} + \beta_3 CD_{i(t-k)} + \beta_4 \langle KD \rangle_{i(t-k)} \\ + \beta_5 CC_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it} \tag{13}$$

These covariates include previous success dummy $PrevSucc_{i(t-k)}$, cumulative trials conducted $Trials_{i(t-k)}$, collaboration diversity $CD_{i(t-k)}$, mean knowledge distance $\langle KD \rangle_{i(t-k)}$, and local clustering coefficient $CC_{i(t-k)}$.

The 3 sets of models for each time lag (1 year, 2 years, and 5 years) resulted in a total of 9 models with results in Tables S6 - S8. We also included a comparison of these models using a likelihood ratio test in Table S9 since the three models are nested within each other. The results show that most of our selected variable's P-values in the main model for all three lag sizes are significant to at least $P < .01$ with most of them being significant to $P < .001$. All VIFs are less than 3 for all models. Previous success has the largest positive effect and that effect size was decreased with as lag sizes increased. Mean knowledge distance has a negative effect for 1 and 2 years lags, while the local clustering coefficient has a negative effect on cumulative trial success for all lag sizes.

As with many types of count data, there's significant overdispersion in the response variable, CTS_{it} – this is supported by the index of dispersion Cox (1966), in which CTS_{it} has an index of dispersion equal to 34.403. The negative-binomial regression handles this overdispersion with a θ shape parameter that captures random effects from unobserved heterogeneity.

S4.2 Trial Success Rate Regression

The trial success rate was analyzed using a beta regression for lags of 1, 2, and 5 years. Trials success rate SR_{it} is defined as the cumulative number of successful clinical trials divided by the total number of trials conducted up to time t for actor i. Trial success rate measures the research efficiency of an actor.

$$SR_{it} = \frac{CTS_{it}}{Trials_{it}} \tag{14}$$

Since the trials success rate in our panel dataset includes zero and one values, which are not included in the (0,1) domain that a beta regression is used to predict, we transformed the SR

with the function, $\frac{y\cdot(n-1)+k}{n}$ as suggested by Zeileis et al. (2010); Smithson and Verkuilen (2006) so that the response variable is between zero and one. The comprehensive model with all measured variables (refer to Section S3) is presented in the main article and mathematically defined as

Comprehensive model (All Variables):

$$\begin{aligned}\text{logit}(SR_{it}) = \beta_0 &+ \beta_1 PrevExp_{i(t-k)} + \beta_2\langle KD\rangle_{i(t-k)} + \beta_3 CC_{i(t-k)} + \beta_4 RD_{i(t-k)} \\ &+ \beta_5\langle RD\rangle_{i(t-k)} + \beta_6 BT_{i(t-k)} + \beta_7 CD_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it} \qquad (15)\end{aligned}$$

For all models predicting trial success rate, the continuous variables, mean knowledge distance $\langle KD\rangle_{it}$, clustering coefficient CC_{it}, research diversification RD_{it}, mean neighbor's research diversification $\langle RD\rangle_{it}$, betweenness centrality BT_{it}, and collaboration diversity CD_{it} were standardized such that the mean is zero and the standard deviation is one. The dummy variable $PrevExp_{it}$ is meant to capture whether the organization has previous experience in clinical trials. Specifically, $PrevExp_{it}$ is equal to 1 if the actor has conducted more than 6 trials[3] before time $t-k$. Since the data is nonstationary, we controlled for time using fixed-effect dummy variables γ_t. The fixed effects for each organizational class (defined in Table S2) are captured in κ_i. For our beta regression, we assumed the link function is an identity function per convention. The coefficient estimates for all comprehensive models with respect to lags are listed as Model B1-3, B2-3, and B3-3 in Tables S10 - S12.

Similar to the analysis of cumulative trial successes in Section S4.1, we developed two sets of models using a beta regression with only control variables and selected variables to show robustness of our comprehensive model. The base beta regression models with only the control variables are listed as Models B1-1, B2-1, and B3-1 in Tables S10 - S12. The reduced models are listed as Models B1-2, B2-2, and B3-2 in Tables S10 - S12 which include selected measured variables that were heuristically determined to maximize predictive accuracy in a random holdout test similar to our analysis in the previous section S4.1 using a negative-binomial regression on cumulative trial

[3] we used 6 because it is the average number of trials

successes. We explicitly show the base and reduced models in (16) and (17), respectively.

Base model (Control Variables):

$$\text{logit}(SR_{it}) = \beta_0 + \beta_1 PrevExp_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it} \tag{16}$$

Reduced model (Selected Variables):

$$\begin{aligned}\text{logit}(SR_{it}) = \beta_0 + \beta_1 PrevExp_{i(t-k)} + \beta_2 \langle KD \rangle_{i(t-k)} + \beta_3 CC_{i(t-k)} + \beta_4 RD_{i(t-k)} \\ + \beta_5 \langle RD \rangle_{i(t-k)} + \gamma_t + \kappa_i + \epsilon_{it}\end{aligned} \tag{17}$$

The regression results for all nine models are shown in Tables S10 - S12. In our base model, most variables with significance of least $P < .001$. The previous experience dummy variable was not significant for a 5-year lag. The local clustering coefficient was the only variable with a negative effect on all lag sizes.

S5 Supplementary Tables and Figures

Table S4: Summary Statistics of Variables (Unstandardized)

Statistic	N	Mean	St. Dev.	Min	Max
Cumulative Trial Success	9,775	2.684	9.610	0	187
Trial Success Rate	9,775	0.283	0.369	0	1.000
Cumulative Trials Conducted	9,775	5.451	19.044	0	398
Mean Knowledge Distance	9,775	0.768	0.272	0	1.414
Local Clustering Coefficient	9,775	0.441	0.294	0	1.000
Vertex Degree	9,775	27.673	39.929	2	258
Research Diversification	9,775	0.334	0.536	0	2.262
Mean Neighbor Research Diversification	9,775	0.706	0.438	0	2.262
Betweenness Centrality	9,775	4,815.421	9,910.420	0	110,661.500
Collaboration Diversity	9,775	1.100	0.676	0	2.344

Table S5: Summary of Variables

Variable	Description	Units	Source
Dependent Variables			
Cumulative Trial Success	Number of trials (phases) conducted up to time t that led to an FDA approved treatment	trials	BiomedTracker
Trial Success Rate	Rate of trials (phases) that led to an FDA approved treatment at time t	%	BiomedTracker
Independent Variables (Lagged)			
Mean Knowledge Distance	The mean of all incident edge's (i.e. collaborations) knowledge distances to the actor	n/a	AACT
Local Clustering Coefficient	Measures the degree to which a node's neighbors are a clique	n/a	AACT
Vertex Degree	The number edge's (collaborations) an actor is involved with	links	AACT
Research Diversification	Entropic measure of research mix (portfolio)	n/a	AACT/BiomedTracker
Mean Neighbor Research Diversification	Mean value of all neighbor collaborator's research diversity	n/a	AACT/BiomedTracker
Betweenness Centrality	Network measure determining the extent that a node is a bridge	n/a	AACT
Cumulative Trials Conducted	The number of trials that an actor has conducted up up to time t	trials	AACT
Collaboration Diversity	The diversity of collaborators with respect to their specialization	n/a	AACT/BiomedTracker
Previous Success	Actor has participated in at least one trial that led to an FDA approved treatment	0/1	BiomedTracker
Previous Experience	Actor has experience in conducting at least six trials	0/1	AACT

Table S6: Negative Binomial Regression on Cumulative Trials Success (1-year lag)

	Cumulative Trials Success		
	(Model A1-1)	(Model A1-2)	(Model A1-3)
	Control Variables	Selected Variables	All Variables
Previous Success	2.665***	2.709***	2.572***
	(0.036)	(0.036)	(0.038)
Cumulative Trials Conducted	0.479***	0.331***	0.290***
	(0.008)	(0.007)	(0.009)
Collaboration Diversity		0.240***	0.182***
		(0.015)	(0.017)
Mean Knowledge Distance		−0.330***	−0.248***
		(0.016)	(0.018)
Local Clustering Coef.		−0.075***	−0.064***
		(0.014)	(0.014)
Mean Neighbor Research Diversification			0.041***
			(0.013)
Betweenness Centrality			0.015
			(0.011)
Research Diversification			0.135***
			(0.015)
Constant	−2.317***	−2.623***	−2.460***
	(0.135)	(0.130)	(0.130)
Observations	9,775	9,775	9,775
Log Likelihood	−11,613.780	−11,256.750	−11,212.030
θ	2.651*** (0.090)	3.419*** (0.131)	3.635*** (0.143)
Akaike Inf. Crit.	23,279.560	22,571.500	22,488.050

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table S7: Negative Binomial Regression on Cumulative Trials Success (2-year lag)

	Cumulative Trials Success		
	(Model A2-1)	(Model A2-2)	(Model A2-3)
	Control Variables	Selected Variables	All Variables
Previous Success	2.113^{***}	2.177^{***}	2.029^{***}
	(0.030)	(0.031)	(0.035)
Cumulative Trials Conducted	0.438^{***}	0.301^{***}	0.251^{***}
	(0.009)	(0.008)	(0.010)
Collaboration Diversity		0.270^{***}	0.213^{***}
		(0.015)	(0.017)
Mean Knowledge Distance		-0.263^{***}	-0.194^{***}
		(0.017)	(0.018)
Local Clustering Coef.		-0.084^{***}	-0.074^{***}
		(0.014)	(0.014)
Mean Neighbor Research Diversification			0.056^{***}
			(0.014)
Betweenness Centrality			0.026^{**}
			(0.013)
Research Diversification			0.128^{***}
			(0.015)
Constant	-1.531^{***}	-1.877^{***}	-1.723^{***}
	(0.116)	(0.111)	(0.112)
Observations	9,131	9,131	9,131
Log Likelihood	−11,870.990	−11,578.300	−11,536.560
θ	2.230^{***} (0.081)	2.850^{***} (0.116)	2.987^{***} (0.125)
Akaike Inf. Crit.	23,789.990	23,210.600	23,133.110

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table S8: Negative Binomial Regression on Cumulative Trials Success (5-year lag)

	Cumulative Trials Success		
	(Model A3-1)	(Model A3-2)	(Model A3-3)
	Control Variables	Selected Variables	All Variables
Previous Success	1.390***	1.285***	1.127***
	(0.044)	(0.046)	(0.051)
Cumulative Trials Conducted	0.321***	0.236***	0.092***
	(0.015)	(0.015)	(0.021)
Collaboration Diversity		0.269***	0.186***
		(0.021)	(0.022)
Mean Knowledge Distance		0.045**	0.092***
		(0.020)	(0.021)
Local Clustering Coef.		−0.154***	−0.128***
		(0.019)	(0.019)
Mean Neighbor Research Diversification			0.068***
			(0.018)
Betweenness Centrality			0.125***
			(0.022)
Research Diversification			0.159***
			(0.022)
Constant	−0.692***	−0.797***	−0.743***
	(0.093)	(0.092)	(0.092)
Observations	6,094	6,094	6,094
Log Likelihood	−9,527.913	−9,378.134	−9,335.782
θ	1.115*** (0.043)	1.289*** (0.053)	1.338*** (0.056)
Akaike Inf. Crit.	19,091.830	18,798.270	18,719.560

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table S9: Likelihood Ratio Test for Cumulative Trials Success Regression Models

Lag	Model	theta	Resid. df	2 x log-lik.	Test	df	LR stat.	Pr.Chi.
1 year	A1-1	2.65	9749	-23225.56				
1 year	A1-2	3.42	9746	-22511.50	A1-1 vs A1-2	3	714.06	0.00
1 year	A1-3	3.63	9743	-22422.05	A1-2 vs A1-3	3	89.45	0.00
2 year	A2-1	2.23	9107	-23739.99				
2 year	A2-2	2.85	9104	-23154.60	A2-1 vs A2-2	3	585.39	0.00
2 year	A2-3	2.99	9101	-23071.11	A2-2 vs A2-3	3	83.49	0.00
5 year	A3-1	1.12	6076	-19053.83				
5 year	A3-2	1.29	6073	-18754.27	A3-1 vs A3-2	3	299.56	0.00
5 year	A3-3	1.34	6070	-18669.56	A3-2 vs A3-3	3	84.70	0.00

Table S10: Beta Regression on Trials Success Rate (1-year lag)

	Trials Success Rate		
	(Model B1-1)	(Model B1-2)	(Model B1-3)
	Control Variables	Selected Variables	All Variables
Previous Experience	0.404^{***}	0.236^{***}	0.238^{***}
	(0.043)	(0.052)	(0.053)
Mean Knowledge Distance		0.025^{*}	0.028^{**}
		(0.014)	(0.014)
Local Clustering Coef.		-0.037^{***}	-0.041^{***}
		(0.014)	(0.015)
Research Diversification		0.092^{***}	0.097^{***}
		(0.019)	(0.020)
Mean Neighbor Research Diversification		0.100^{***}	0.102^{***}
		(0.014)	(0.014)
Betweenness Centrality			0.0003
			(0.018)
Collaboration Diversity			−0.020
			(0.017)
Constant	-0.533^{***}	-0.434^{***}	-0.422^{***}
	(0.146)	(0.147)	(0.147)
Observations	9,775	9,775	9,775
R^2	0.037	0.052	0.053
Log Likelihood	36,126.930	36,171.430	36,172.170

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table S11: Beta Regression on Trials Success Rate (2-year lag)

	Trials Success Rate		
	(Model B2-1)	(Model B2-2)	(Model B2-3)
	Control Variables	Selected Variables	All Variables
Previous Experience	0.412***	0.196***	0.196***
	(0.049)	(0.061)	(0.061)
Mean Knowledge Distance		0.034**	0.038***
		(0.014)	(0.014)
Local Clustering Coef.		−0.041***	−0.044***
		(0.015)	(0.015)
Research Diversification		0.103***	0.109***
		(0.020)	(0.021)
Mean Neighbor Research Diversification		0.115***	0.118***
		(0.014)	(0.014)
Betweenness Centrality			0.006
			(0.019)
Collaboration Diversity			−0.029
			(0.018)
Constant	−0.787***	−0.671***	−0.658***
	(0.116)	(0.117)	(0.118)
Observations	9,131	9,131	9,131
R^2	0.037	0.057	0.057
Log Likelihood	32,728.370	32,782.770	32,784.110

*p<0.1; **p<0.05; ***p<0.01

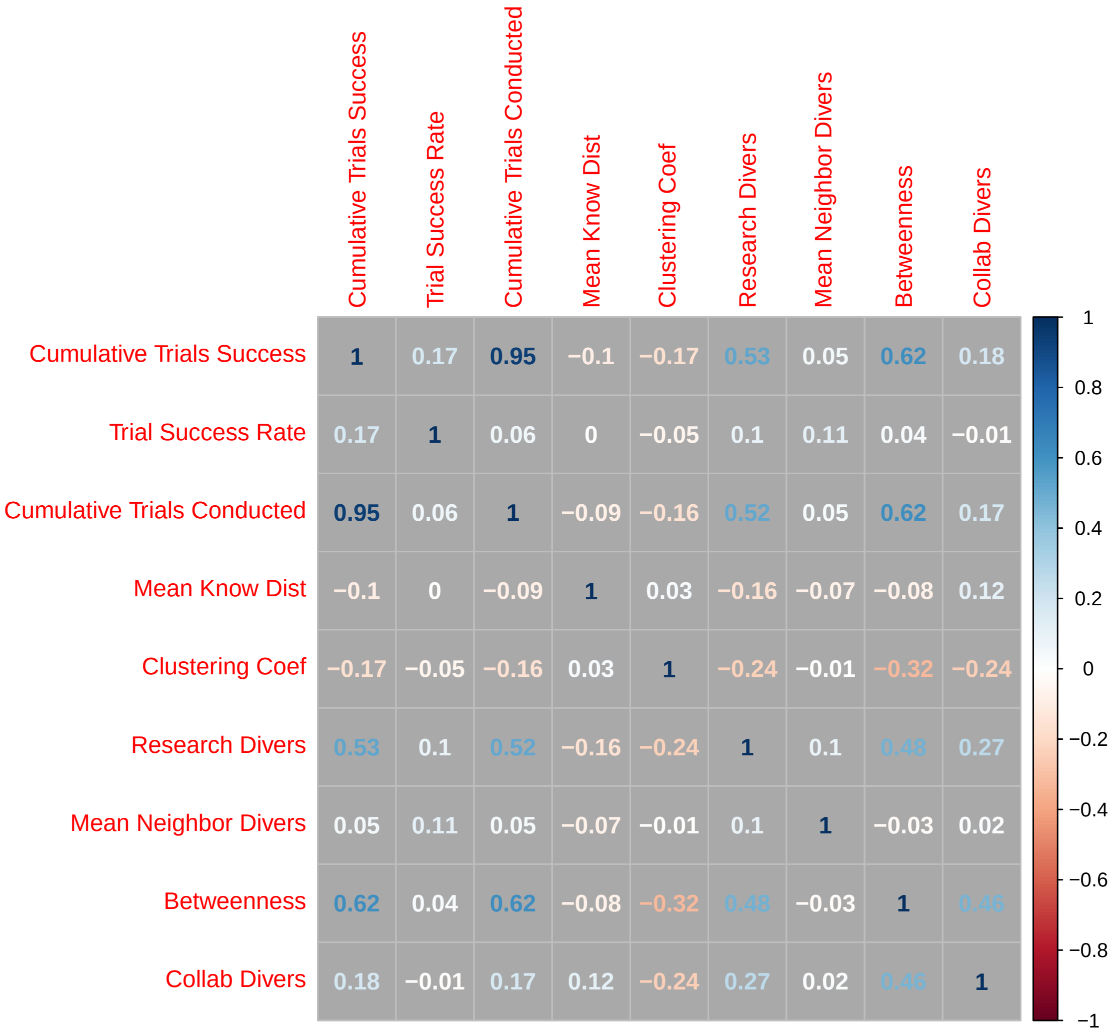
Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.

Table S12: Beta Regression on Trials Success Rate (5-year lag)

	Trials Success Rate		
	(Model B3-1)	(Model B3-2)	(Model B3-3)
	Control Variables	Selected Variables	All Variables
Previous Experience	0.362***	0.077	0.055
	(0.083)	(0.105)	(0.107)
Mean Knowledge Distance		0.031*	0.038**
		(0.018)	(0.018)
Local Clustering Coef.		−0.040**	−0.041**
		(0.018)	(0.018)
Research Diversification		0.102***	0.100***
		(0.026)	(0.027)
Mean Neighbor Research Diversification		0.110***	0.113***
		(0.017)	(0.018)
Betweenness Centrality			0.033
			(0.024)
Collaboration Diversity			−0.040*
			(0.021)
Constant	−0.822***	−0.750***	−0.740***
	(0.087)	(0.088)	(0.089)
Observations	6,094	6,094	6,094
R^2	0.033	0.050	0.051
Log Likelihood	18,953.220	18,986.650	18,988.750

*p<0.1; **p<0.05; ***p<0.01

Note: The fixed effects of time and organizational class are not shown. All underlined variables are standardized.



	Cumulative Trials Success	Trial Success Rate	Cumulative Trials Conducted	Mean Know Dist	Clustering Coef	Research Divers	Mean Neighbor Divers	Betweenness	Collab Divers
Cumulative Trials Success	1	0.17	0.95	−0.1	−0.17	0.53	0.05	0.62	0.18
Trial Success Rate	0.17	1	0.06	0	−0.05	0.1	0.11	0.04	−0.01
Cumulative Trials Conducted	0.95	0.06	1	−0.09	−0.16	0.52	0.05	0.62	0.17
Mean Know Dist	−0.1	0	−0.09	1	0.03	−0.16	−0.07	−0.08	0.12
Clustering Coef	−0.17	−0.05	−0.16	0.03	1	−0.24	−0.01	−0.32	−0.24
Research Divers	0.53	0.1	0.52	−0.16	−0.24	1	0.1	0.48	0.27
Mean Neighbor Divers	0.05	0.11	0.05	−0.07	−0.01	0.1	1	−0.03	0.02
Betweenness	0.62	0.04	0.62	−0.08	−0.32	0.48	−0.03	1	0.46
Collab Divers	0.18	−0.01	0.17	0.12	−0.24	0.27	0.02	0.46	1

Figure S10: Correlation matrix for the network, organizational, and collaboration measures over the 2006-2016 time frame at 6-month intervals with one year lag for Cumulative Trial Success and Trial Success Rate. (N = 9055).

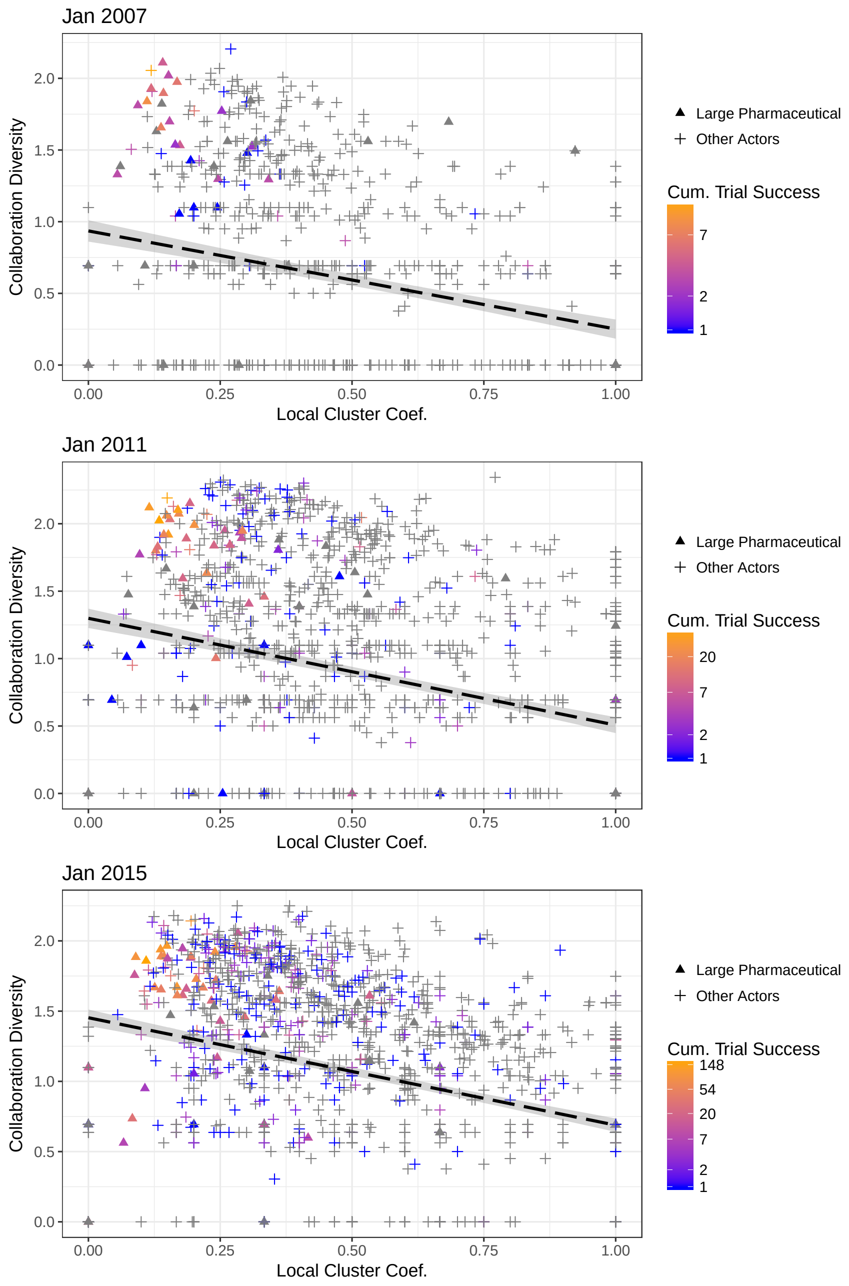


Figure S11: The scatterplot comparing collaboration diversity and local clustering coefficient for each actor. Large pharamceutical companies and other actors are distinguished by shapes. The color gradient is scaled based on cumulative trial successes.

References

Bar, T. and Leiponen, A. (2012). A measure of technological distance. *Economics Letters*, 116(3):457–459.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177.

Cockburn, I. M. and Henderson, R. M. (2001). Scale and scope in drug development: unpacking the advantages of size in pharmaceutical research. *Journal of health economics*, 20(6):1033–1057.

Cox, D. R. (1966). The statistical analysis of series of events. *Monographs on Applied Probability and Statistics.*

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. *Quality and Quantity*, 14(4):585–592.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Guler, I. and Nerkar, A. (2012). The impact of global and local cohesion on innovation in the pharmaceutical industry. *Strategic Management Journal*, 33(5):535–549.

Makri, M., Hitt, M. A., and Lane, P. J. (2010). Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. *Strategic Management Journal*, 31(6):602–628.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Schilling, M. A. and Phelps, C. C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management Science*, 53(7):1113–1126.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54.

Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B. J., and Pietrobon, R. (2012). The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty. *PloS One*, 7(3):e33677.

Thomas, D. W., Burns, J., Audette, J., Carroll, A., Dow-Hygelund, C., and Hay, M. (2016). Clinical development success rates 2006–2015. *San Diego: Biomedtracker/Washington, DC: BIO/Bend: Amplion.*

Vaccario, G., Tomasello, M. V., Tessone, C. J., and Schweitzer, F. (2018). Quantifying knowledge exchange in r&d networks: A data-driven model. *Journal of Evolutionary Economics*, 28(3):461–493.

Zeileis, A., Cribari-Neto, F., Grün, B., and Kos-midis, I. (2010). Beta regression in r. *Journal of Statistical Software*, 34(2):1–24.