UNIVERSITÉ PARIS DESCARTES

École Doctorale Pierre Louis de Santé Publique:
Épidémiologie et Sciences de l'Information Biomédicale

Centre de Recherche Épidémiologie et Statistique Sorbonne Paris Cité,
INSERM UMR1153, équipe METHODS

Modélisation du système complexe de la publication scientifique

Michail Kovanis

Thèse de doctorat de Santé Publique
Spécialité Épidémiologie

Dirigée par Raphaël Porcher
Co-encadrée par Ludovic Trinquart

Présentée et soutenue publiquement le 2 Octobre 2017

Devant un jury composé de:

Rapporteurs:	
Geneviève CHENE	[Professeur] - Université de Bordeaux
Didier TORNY	[Directeur de recherche] - Centre National de Recherche Scientifique
Examinateurs:	
Philippe RAVAUD	[Professeur] - Université Paris Descartes
Flaminio SQUAZZONI	[Maître de conférences] - University of Brescia
Dominique COSTAGLIOLA	[Directrice de recherche] - Université Pierre et Marie Curie
Directeur de thèse:	
Raphaël PORCHER	[Maître de conférences] - Université Paris Descartes

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Dr. Raphaël Porcher and my co-advisor Dr. Ludovic Trinquart for all the patience and support during this PhD thesis. Their mentorship was crucial for keeping me motivated and to successfully complete this project.

Besides my advisors, I would like to express my gratitude to Prof. Philippe Ravaud for his guidance and for making all this possible. Moreover, I thank the reviewers, Prof. Geneviève Chêne and Dr. Didier Torny, and the examiners Dr. Dominique Costagliola and Dr. Flaminio Squazzoni for not hesitating to accept my invitation to provide their very useful insights regarding my work in this thesis.

I thank my friends for being there when I mostly needed them. Especially, I thank Martin for all the deep conversations that inspired me and motivated me to become better. I would like to thank Ignacio for his friendship, and his moral and technical support throughout all the years of my PhD thesis. I also thank Alex for his great help on copyediting the manuscript.

Last but not least, I would like to thank my parents and my brother for supporting me through all the difficulties in my life and my studies.

Affiliations

1. INSERM U1153, Paris, France

2. Université Paris Descartes–Sorbonne Paris cité, Paris, France

Funding

Université Paris Descartes - Sorbonne Paris Cité, Paris, France

List of publications

Paper 1: Kovanis M, Porcher R, Ravaud P, Trinquart L (2016) The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise. PLoS ONE 11(11): e0166387. doi: `http://dx.doi.org/10.1371/journal.pone.0166387`.

Paper 2: Kovanis, M., Porcher, R., Ravaud, P., Trinquart, L. Complex systems approach to scientific publication and peer-review system: development of an agent-based model calibrated with empirical journal data. Scientometrics (2016) 106: 695. doi: `http://dx.doi.org/10.1007/s11192-015-1800-6`.

Paper 3: Kovanis, M., Trinquart, L., Ravaud, P., Porcher, R. Evaluating alternative systems of peer review: a large-scale agent-based modelling approach to scientific publication. Scientometrics (2017). doi: `http://dx.doi.org/10.1007/s11192-017-2375-1`.

Résumé: Le système d'évaluation par les pairs est le *gold standard* de la publication scientifique. Ce système a deux objectifs: d'une part filtrer les articles scientifiques erronés ou non pertinents et d'autre part améliorer la qualité de ceux jugés dignes de publication. Le rôle des revues scientifiques et des rédacteurs en chef est de veiller à ce que des connaissances scientifiques valides soient diffusées auprès des scientifiques concernés et du public. Cependant, le système d'évaluation par les pairs a récemment été critiqué comme étant intenable sur le long terme, inefficace et cause de délais de publication des résultats scientifiques. Dans ce projet de doctorat, j'ai utilisé une modélisation par systèmes complexes pour étudier le comportement macroscopique des systèmes de publication et d'évaluation par les pairs. Dans un premier projet, j'ai modélisé des données empiriques provenant de diverses sources comme Pubmed et Publons pour évaluer la viabilité du système. Je montre que l'offre dépasse de 15% à 249% la demande d'évaluation par les pairs et, par conséquent, le système est durable en termes de volume. Cependant, 20% des chercheurs effectuent 69% à 94% des revues d'articles, ce qui souligne un déséquilibre significatif en termes d'efforts de la communauté scientifique. Les résultats ont permis de réfuter la croyance largement répandue selon laquelle la demande d'évaluation par les pairs dépasse largement l'offre mais ont montré que la majorité des chercheurs ne contribue pas réellement au processus. Dans mon deuxième projet, j'ai développé un modèle par agents à grande échelle qui imite le comportement du système classique d'évaluation par les pairs, et que j'ai calibré avec des données empiriques du domaine biomédical. En utilisant ce modèle comme base pour mon troisième projet, j'ai modélisé cinq systèmes alternatifs d'évaluation par les pairs et évalué leurs performances par rapport au système conventionnel en termes d'efficacité de la revue, de temps passé à évaluer des manuscrits et de diffusion de l'information scientifique. Dans mes simulations, les deux systèmes alternatifs dans lesquels les scientifiques partagent les commentaires sur leurs manuscrits rejetés avec les éditeurs du prochain journal auquel ils les soumettent ont des performances similaires au système classique en termes d'efficacité de la revue. Le temps total consacré par la communauté scientifique à l'évaluation des articles est cependant réduit d'environ 63%. En ce qui concerne la dissémination scientifique, le temps total de la première soumission jusqu'à la publication est diminué d'environ 47% et ces systèmes permettent de diffuser entre 10% et 36% plus d'informations scientifiques que le système conventionnel. Enfin, le modèle par agents développé peut être utilisé pour simuler d'autres systèmes d'évaluation par les pairs ou des interventions, pour ainsi déterminer les interventions ou modifications les plus prometteuses qui pourraient être ensuite testées par des études expérimentales en vie réelle.

Mots-clés: Evaluation par les pairs; Systèmes complexes; Modélisation par agents

English title: Modeling the complex system of scientific publication

Abstract: The peer-review system is undoubtedly the gold standard of scientific publication. Peer review serves a two-fold purpose; to screen out of publication articles containing incorrect or irrelevant science and to improve the quality of the ones deemed suitable for publication. Moreover, the role of the scientific journals and editors is to ensure that valid scientific knowledge is disseminated to the appropriate target group of scientists and to the public. However, the peer-review system has recently been criticized, in that it is unsustainable, inefficient and slows down publication. In this PhD thesis, I used complex-systems modeling to study the macroscopic behavior of the scientific publication and peer-review systems. In my first project, I modeled empirical data from various sources, such as Pubmed and Publons, to assess the sustainability of the system. I showed that the potential supply has been exceeding the demand for peer review by 15% to 249% and thus, the system is sustainable in terms of volume. However, 20% of researchers have been performing 69% to 94% of the annual reviews, which emphasizes a significant imbalance in terms of effort by the scientific community. The results provided evidence contrary to the widely-adopted, but untested belief, that the demand for peer review over-exceeds the supply, and they indicated that the majority of researchers do not contribute to the process. In my second project, I developed a large-scale agent-based model, which mimicked the behavior of the conventional peer-review system. This model was calibrated with empirical data from the biomedical domain. Using this model as a base for my third project, I developed and assessed the performance of five alternative peer-review systems by measuring peer-review efficiency, reviewer effort and scientific dissemination as compared to the conventional system. In my simulations, two alternative systems, in which scientists shared past reviews of their rejected manuscripts with the editors of the next journal to which they submitted, performed equally or sometimes better in terms of peer-review efficiency. They also each reduced the overall reviewer effort by ∼63%. In terms of scientific dissemination, they decreased the median time from first submission until publication by ∼47% and diffused on average 10% to 36% more scientific information (i.e., manuscript intrinsic quality x journal impact factor) than the conventional system. Finally, my agent-based model may be an approach to simulate alternative peer-review systems (or interventions), find those that are the most promising and aid decisions about which systems may be introduced into real-world trials.

Keywords: Peer review, Complex systems, Agent-based modeling

Synthèse des travaux de thèse

Objectifs

Les objectifs de mon projet de doctorat étaient (i) d'identifier le fardeau que fait peser l'évaluation par les pairs sur la communauté scientifique, (ii) de développer une méthode de simulation du système de publication scientifique conventionnelle et d'évaluation par les pairs et (iii) d'utiliser ce cadre de simulations pour comparer l'efficacité du système conventionnel avec celle de systèmes alternatifs.

Le système des publications scientifiques

La science est la meilleure méthode pour acquérir des connaissances fiables. Le système de publication scientifique est censé être à la fois le gardien et le vecteur de la diffusion des découvertes scientifiques. La publication scientifique repose en grande partie sur un système d'évaluation par les pairs, au cours duquel une communication scientifique (article) est évaluée par d'autres chercheurs (pairs) avant d'être publiée. Ceci permet de s'assurer que la science peu pertinente ou mal conduite ne soit pas publiée, tout en aidant à améliorer la qualité des manuscrits jugés comme étant publiables (Rennie 2003; Sense About Science 2004). D'autre part, les revues scientifiques et les éditeurs sont responsables de s'assurer que toutes les connaissances scientifiques valides soient publiées et atteignent le groupe cible approprié dans la communauté scientifique. Les acteurs les plus importants du système de publication scientifique sont donc les chercheurs, les revues scientifiques et les éditeurs, qui interagissent tous les uns avec les autres.

La publication scientifique a une longue histoire. La première revue consacrée exclusivement à la science était *Philosophical Transactions of the Royal Society*, qui a été publiée pour la première fois en 1665 (Spier 2002). À cette époque, la revue publiait uniquement des articles choisis par son éditeur, Henry Oldenburg, et ne passaient pas par une quelconque évaluation par les pairs. L'édition dans les revues scientifiques a commencé à prendre une forme plus proche de sa forme actuelle durant le milieu du 19$^{\text{ème}}$ siècle (Rennie 2003). À l'heure actuelle, la publication scientifique est un système très complexe avec un grand nombre d'interactions entre de multiples agents hétérogènes. En outre, ces dernières décennies ont vu la taille du système augmenter rapidement. Par exemple, on a estimé, qu'en 2006, il y avait environ 1,35 million de publications publiées dans environ 24 000 revues (Björk 2008). Neuf ans plus tard, le nombre de publications a presque doublé (2,5 millions) et le nombre de revues scientifiques est passé à 28 000 (Ware et Mabe 2015).

Même si l'édition scientifique est ancienne, l'évaluation par les pairs en tant que système est beaucoup plus récente que ce que la plupart des gens croiraient. Par exemple, *Nature* a introduit un processus officiel d'évaluation par les pairs en 1967 et le *Lancet* en 1976. Avant cela, les scientifiques n'étaient pas familiers avec l'évaluation par les pairs. Par exemple, Einstein et Rosen ont réagit vivement à l'envoi de leur article soumis à *Physical Review* pour une évaluation par les pairs, parce qu'ils n'avaient pas autorisé la revue à partager leurs recherches avec d'autres scientifiques avant leur publication (Rennie 2003; Csiszar 2016).

Traditionnellement, les revues scientifiques sont accessibles aux lecteurs, aux institutions, ou aux bibliothèques, entre autres, moyennant un abonnement. Cependant, après le développement d'internet dans les premières années du 21$^{\text{ème}}$ siècle, cette manière conventionnelle d'édition a été contestée. L'émergence de revues en ligne et de dépôts en ligne a introduit un nouveau mode de publication gratuit pour les lecteurs, mais pas nécessairement pour les auteurs: l'accès ouvert (*open access*). L'accès ouvert est une autre forme d'édition, dans laquelle les auteurs (ou leurs financeurs) peuvent payer des frais à une revue scientifique pour que leur article soit publié et accessible librement sur le site Web de celle-ci (*gold open access*), ou bien peuvent téléverser leur article sur des serveurs de pré-publication (*green open access*). Les frais du *gold open access* correspondent, dans la majorité des cas, à des frais d'édition ou d'abonnement. Les revues peuvent renoncer à ces frais dans le cas où les auteurs n'ont pas les moyens de les payer (auteurs dans des pays en voie de développement, par exemple). Les manuscrits publiés via le *green open access* ne passent pas par une évaluation par les pairs avant d'être téléversés sur un serveur, mais généralement sont aussi soumis pour publication par les auteurs à des revues scientifiques, dans lesquelles ils seront l'objet d'une évaluation par les pairs traditionnelle. De nos jours, l'accès ouvert domine la publication scientifique, et même les revues dites "traditionnelles" au format papier proposent aux auteurs une option pour une publication en *gold open access* (Bohannon 2014).

La culture de l'évaluation par les pairs et de la publication scientifique peut différer en fonction du domaine scientifique. Par exemple, en physique et en mathématiques, il est courant que des articles soient d'abord pré-publiés sur un serveur en ligne de l'Université Cornell (ArXiv) avant d'être soumis à une revue scientifique et de subir une quelconque évaluation traditionnelle par les pairs. Ainsi, de nombreux articles d'ArXiV sont partagés et discutés par la communauté dans des forums ou des médias sociaux en ligne avant d'être publiés par une revue. Ce modèle a aussi été adopté en biologie (via bioRxiV), économie (RePEc) et d'autres domaines (PhilSci-Archive, PsyArXiv, ChemRxiv, MedArXiv etc.). Toutefois, dans la plupart des domaines scientifiques, l'habitude de discuter les articles avant l'examen formel par les pairs est beaucoup moins répandue qu'en physique et mathématiques.

Il existe plusieurs métriques pour mesurer l'impact relatif des revues scientifiques: le facteur d'impact (*impact factor*), le *eigenfactor*, la demi-vie de citation, entre autres. Le plus célèbre est le facteur d'impact d'une revue. Il est défini comme le nombre moyen de citations que les articles d'une revue a reçu sur une période de deux ans (Garfield 2006; Alberts 2013). Selon le domaine scientifique, les auteurs peuvent essayer de publier leurs articles dans les revues au facteur d'impact le plus élevé possible en fonction de l'importance perçue de leur manuscrit.

Édition et processus d'évaluation par les pairs

Lorsque les scientifiques finissent un travail de recherche, ils doivent résumer et présenter leurs résultats dans un rapport. Ce rapport doit ensuite être communiqué au public approprié afin de mettre à jour les connaissances scientifiques de toutes les parties intéressées. Traditionnellement, la diffusion des rapports scientifiques ou des articles est assurée par les revues scientifiques, qui publient des travaux de recherche sur un sujet spécifique ou général. Les auteurs choisissent de soumettre leurs travaux à des revues qui maximiseront leur public, et les revues doivent aussi sélectionner des articles qui maximiseront l'intérêt et la taille de leur public. Ainsi, les revues scientifiques mettent en œuvre des techniques d'évaluation pour s'assurer que le contenu qu'ils publient est valide, pertinent et intéressant pour leur public.

Les manuscrits soumis passent par un processus d'évaluation interne, au cours duquel un éditeur est affecté à chaque manuscrit afin de décider si le manuscrit est pertinent pour la revue. Cette première décision prend habituellement entre quelques heures et quelques jours. Si un manuscrit est rejeté, il peut être renvoyé à une autre revue. S'il n'est pas rejeté, l'éditeur contacte d'autres scientifiques qui sont considérés comme des experts du domaine ou qui ont déjà publié des articles sur le même sujet. Les invitations envoyées par les éditeurs contiennent habituellement le résumé du manuscrit soumis et, en fonction de cela, les scientifiques invités doivent décider s'ils acceptent de l'évaluer. Les relecteurs candidats peuvent refuser d'évaluer un manuscrit pour diverses raisons, par exemple parce qu'ils manquent de temps, ou bien parce que le manuscrit n'est pas dans leur domaine d'expertise (Mulligan et al. 2013). Ceux qui acceptent d'évaluer le manuscrit le font généralement en tant que bénévoles, même si certaines revues peuvent choisir de les récompenser avec, par exemple, des bons de réduction sur des frais de publication. En outre, certaines revues publient une liste annuelle des noms de leurs relecteurs en guise de remerciement. Il existe aussi des plateformes de reconnaissance en ligne comme Publons qui permettent d'afficher cette part du travail des scientifiques (Review rewards 2014; Warne 2016). L'évaluation par les pairs est habituellement conduite en simple ou double aveugle, ce qui signifie que les auteurs ne connaissent pas l'identité des relecteurs, alors que ces derniers connaissent (simple aveugle) ou pas (double aveugle) l'identité des auteurs. Cela est mis en place pour minimiser la possibilité de représailles de la part d'auteurs qui auraient reçu des évaluations négatives de leur travail. Néanmoins, il existe des revues qui rendent publiques les noms (et parfois les évaluations) de leurs relecteurs (par exemple BMC et BMJ Open).

Les éditeurs, après avoir obtenu suffisamment de rapports d'évaluation (généralement entre un et trois, parfois plus), prennent une décision quant au rejet du manuscrit, ou bien demandent aux auteurs de faire des modifications au manuscrit en fonction des commentaires des relecteurs et de le soumettre de nouveau pour un deuxième tour d'évaluations. Dans certaines revues, ces décisions sont prises lors d'une réunion périodique du comité de rédaction. Les manuscrits rejetés peuvent être soumis à une autre revue, généralement après quelques modifications. Les manuscrits ayant fait l'objet d'une demande de modifications sont réévalués et révisés autant de fois que nécessaire jusqu'à ce qu'une décision finale de rejet ou d'acceptation soit faite, bien que plus de deux ou trois tours d'évaluation soit rare. Les manuscrits acceptés sont inclus dans un numéro de la revue, peuvent être téléversés sur le site Web de la revue beaucoup plus tôt que la publication imprimée. Selon le domaine scientifique, toute la procédure peut prendre de quelques mois à plus d'un an, alors qu'un seul rapport d'évaluation nécessite habituellement seulement quelques heures de travail de la part du relecteur.

Critique du système d'évaluation par les pairs

L'évaluation par les pairs a récemment été débattue et critiquée (Gura 2002; Smith 2006; Alberts et al. 2008; Stahel et Moore 2014; Rennie 2016; Csiszar 2016). L'augmentation importante du nombre de manuscrits scientifiques a entraîné une augmentation de la demande d'évaluations par les pairs, et a potentiellement introduit un fardeau important pour la communauté scientifique avec un risque de dégradation de la qualité des évaluations. Il a été suggéré que, dans l'ensemble du domaine biomédical, les scientifiques doivent consacrer des dizaines de millions d'heures par an pour effectuer les évaluations par les pairs, dont une grande partie est potentiellement redondante (c'est-à-dire des évaluations multiples pour des manuscrits déjà examinés auparavant) (Kovanis et al. 2016). En outre, la capacité du système à détecter des erreurs dans les manuscrits a également été contestée. Par exemple, un essai contrôlé randomisé conduit par le *British Medical Journal* a montré que les relecteurs pouvaient repérer en moyenne 2 à 3 des 9 principales erreurs méthodologiques dans un manuscrit, même après une formation spécifique (Schroter et al. 2004). Une autre étude a montré que des revues en psychologie peuvent rejeter des articles qu'ils ont déjà publié, lorsqu'ils leur sont re-soumis avec des légères modifications dans le contenu (Peter & Ceci 1982). Enfin, il est estimé que ce système coûte des milliards de dollars aux institutions scientifiques chaque année en raison du temps que les scientifiques y consacrent (Look et Sparks 2010).

Le processus d'évaluation par les pairs peut également être biaisé par des scientifiques ayant des motivations peu éthiques. Un scandale concernant des rapports d'évaluation fabriqués a entraîné de multiples retraits d'articles dans de nombreuses revues (Cat et al. 2014, Callaway 2015, Cohen et al. 2016). L'escroquerie a été révélée lorsque des éditeurs se sont aperçu que certains auteurs proposant des scientifiques renommés en tant que relecteurs de leurs manuscrits fournissaient de fausses adresses électroniques conduisant à eux-mêmes. Ainsi, les auteurs écrivaient des évaluations très favorables sur leurs propres manuscrits, afin de maximiser leur probabilité d'être acceptés. Enfin, un autre scandale impliquant des manuscrits publiés dans des conférences scientifiques a révélé que plus d'une centaine de manuscrits générés automatiquement par un ordinateur avaient été acceptés après leur processus d'évaluation par les pairs (van Noorden 2014).

En outre, après l'adoption de l'accès ouvert par la communauté scientifique, plusieurs revues dites prédatrices sont apparues (Sorokowski et al. 2017). En 2013, le personnel éditorial de *Science* a envoyé un article truqué à de nombreuses revues scientifique proposant le *gold open access* (Bohannon 2013). Même si pour de nombreuses revues le manuscrit a été rejeté après une évaluation éditoriale ou par les pairs, un bon nombre d'entre elles ont accepté le manuscrit presque instantanément après la soumission. Toutes ces revues ont affirmé que le manuscrit était passé par un évaluation par les pairs sans en fournir aucune preuve. Ces journaux ont alors demandé à l'auteur de payer les frais d'accès ouvert, ce qui montre que ces revues cherchent à gagner de l'argent facilement en publiant tous les articles reçus.

Interventions et études sur le système d'évaluation par les pairs

Bien que le processus traditionnel d'évaluation par les pairs soit communément adopté par les revues scientifiques, des modifications à ce processus ont été proposées et mises en œuvre dans certains cas. Par exemple, des revues ont expérimenté la mise en place d'évaluations par les pairs en double aveugle contre ouvert (Blank 1991; van Rooyen et al. 1999; Pöschl 2012; Pontille & Torny 2014; Hopewell 2014; Bruce et al. 2016). D'autres revues ont évalué les effets de formations visant à apprendre aux jeunes relecteurs à évaluer de façon critique des manuscrits scientifiques (Schroter et al. 2004; Houry et al. 2012). Il a aussi été proposé de diviser un manuscrit en différentes parties afin d'être évaluées indépendamment. Par exemple, des relecteurs spécialisés pourraient évaluer la conduite des analyses statistiques, et d'autres juger si les critères de jugement ont été correctement déclarés.

Outre ces interventions à micro-échelle, des modifications macroscopiques ont été proposées et mises en œuvre (Walker et Rocha da Silva 2015). Premièrement, en physique et en mathématiques il a été largement adopté de publier immédiatement les manuscrits avant d'être soumis à une revue. Dans ces domaines, lorsque les auteurs rédigent un manuscrit, ils peuvent le téléverser sur ArXiv et ensuite suivre le processus standard de soumission et d'évaluation par les pairs. Cependant, tous les autres scientifiques peuvent lire le travail, discuter et commenter à ce sujet avant la publication de celui-ci dans une revue. Deuxièmement, certains journaux commencent dernièrement à adopter l'évaluation par les pairs après publication (Herron 2012; Hunter 2012). Ce système alternatif permet à d'autres scientifiques de commenter et d'examiner les manuscrits après avoir été publiés en ligne ou acceptés par une revue. Des journaux mettant actuellement en œuvre ce système sont par exemple *f1000Research*, *eLife*, *Atmospheric Chemistry and Physics* (ACP), entre autres. Un autre système alternatif est celui dans lequel un article, après rejet, est soumis à nouveau dans une autre revue accompagné des évaluations par les pairs de la revue précédente (Cals et al. 2013; van Noorden 2013). Ce système vise à éliminer les examens redondants et à accélérer le processus décisionnel.

Article 1

The global burden of journal peer review in biomedical literature: Strong imbalance in the collective enterprise

La production scientifique a progressivement augmenté au cours des dernières décennies. Une simple requête de la totalité d'articles indexés dans PubMed renvoie respectivement 400 000, 700 000 et 1 200 000 de résultats pour 1996, 2006 et 2016, respectivement. Cela a donc augmenté les besoins en matière d'évaluation par les pairs. Beaucoup de scientifiques éminents ont exprimé leur inquiétude quant au fait que cette augmentation du nombre de publications surcharge les scientifiques et pourrait ne pas être tenable (Arns 2014, Breuning et al. 2015). Étant donné que les articles sont généralement soumis à plus d'une revue avant d'être publiés, et qu'ils sont conséquemment objets de plusieurs évaluations par les pairs, il est facile de penser que ces affirmations peuvent être vraies. Indépendamment de la pertinence de ces allégations, le fardeau global de l'évaluation par les pairs n'a jamais été quantifié de façon précise. Ainsi, nous ne savons vraiment pas si le système est viable ou non. Mon objectif dans cette étude était de quantifier le fardeau général de l'évaluation par les pairs supporté par la communauté scientifique dans le domaine biomédical, et d'estimer comment celui-ci est réparti entre les scientifiques.

J'ai estimé la demande annuelle et l'offre potentielle d'évaluations en utilisant une approche par modélisation mathématique. J'ai utilisé des données issues principalement du domaine biomédical. J'ai téléchargé de Pubmed toute l'information des articles publiés entre 1990 et 2015. De plus, j'ai récupéré les données d'un sondage international fait par Elsevier concernant la resoumission de manuscrits ainsi que le temps passé par les relecteurs dans le processus d'évaluation par les pairs (Mulligan et al. 2013). J'ai dérivé la distribution du nombre d'évaluations de manuscrits effectués par relecteur durant un an à partir des données de Publons.com. Cette base recueille des données provenant de plus de 70 000 relecteurs et 10 000 revues scientifiques. Les variables pour lesquelles aucune donnée n'était disponible, comme par exemple le nombre d'articles non publiés et les taux de rejet éditorial (i.e. sans évaluation par les pairs), des hypothèses et des analyses de sensibilité approfondies ont été effectuées. Enfin, après avoir combiné dans une même équation le taux de rejet éditorial, le nombre moyen de relecteurs par article, le nombre total d'articles soumis, et la probabilité que les articles passent à la deuxième étape d'évaluation par les pairs, j'ai estimé la demande annuelle d'évaluations par les pairs de 1990 à 2015.

Étant donné qu'il n'y a pas de définition communément admise de quel scientifique est qualifié pour évaluer un manuscrit, j'ai utilisé différents scénarios pour estimer le nombre de scientifiques pouvant être relecteurs chaque année. Par exemple, un scénario consisterait à considérer comme possibles relecteurs tous les premiers et derniers co-auteurs d'au moins une publication au cours d'une année. Ce scénario correspondrait à la borne inférieure du nombre de relecteurs disponibles en une année. En effet, les premiers auteurs seraient qualifiés pour être relecteurs en tant qu'investigateurs principaux d'une recherche spécifique, de même que les derniers auteurs, qui eux ont supervisé cette recherche. Une borne supérieure pour le nombre de relecteurs pourrait être dérivée à partir de tous les co-auteurs d'au moins une publication au cours d'une année donnée. La taille réelle de l'ensemble de relecteurs disponibles est susceptible de se situer entre ces deux scénarios. Il est important de noter ici que ces scénarios définissent le nombre de relecteurs potentiels qui seraient disponibles et qualifiés pour examiner les manuscrits, mais non pas le nombre de relecteurs que les éditeurs peuvent contacter ou qui accepteront d'évaluer un manuscrit.

J'ai montré que, pour l'année 2015, dans l'ensemble des scénarios étudiés, l'offre de relecteurs potentiels dépassait la demande de relecteurs nécessités de 15% à 249%.

Cependant, 20% des scientifiques ont effectué 69% à 94% des évaluations par les pairs. Parmi les scientifiques ayant fait au moins une évaluation par les pairs, 70% ont dédié 1% ou moins de leur temps de travail de recherche à cette tâche, tandis que 5% en ont dédié 13% ou plus. On estime que 63,4 millions d'heures ont été consacrées à l'évaluation par les pairs en 2015, dont 18,9 millions d'heures ont été fournies par 5% des relecteurs. Ces résultats montrent que le système est durable en termes de volume, mais soulignent un déséquilibre considérable dans la répartition de l'effort fourni par l'évaluation des manuscrits entre la communauté scientifique. Ainsi, le système n'est pas efficace pour égaliser la charge de travail de l'évaluation par les pairs, et pourrait donc être considéré comme étant insoutenable pour les personnes surchargées par cette tâche.

Modélisation du système d'évaluation par les pairs

Même si l'évaluation par les pairs est une pierre angulaire du système de publication scientifique, ses mécanismes n'ont pas encore été étudiés de manière approfondie. En fait, bien qu'il soit maintenant naturel d'utiliser l'évaluation par les pairs, les mécanismes du ce système ne reposent pas sur des faits. (Rennie 2016; Bruce et al. 2016) Des essais contrôlés randomisés étudiant l'évaluation par les pairs nécessiteraient la participation d'un grand nombre de scientifiques évaluant de nombreux manuscrits dans plusieurs revues. Il serait par ailleurs considérablement compliqué de mettre en place des essais pour évaluer toutes les interventions ou alternatives du système proposées par la communauté scientifique.

Une alternative serait de modéliser l'ensemble du système de publication scientifique et d'évaluation par les pairs, et utiliser des simulations informatiques rapides et peu coûteuses pour étudier les interventions et alternatives au système. Mon objectif dans les deux prochains articles était de créer une simulation par ordinateur du système de publication scientifique et d'évaluation par les pairs.

À cause de la complexité du modèle, j'ai choisi d'utiliser une méthode issue des sciences des systèmes complexes appelée modélisation par agents. Cette méthode permet une modélisation microscopique de chaque agent impliqué dans le système, et permet de dériver des caractéristiques macroscopiques du système résultant de leurs interactions. Cette approche permet de pré-sélectionner des interventions et des alternatives au système, et d'identifier celles qui seraient le plus susceptibles d'améliorer le système actuel. Ainsi, les essais futurs pourraient être guidés par les résultats des simulations afin d'économiser des ressources.

Pour créer un modélisation par agents, il faut d'abord identifier les principales composantes ou agents du système étudié. Les agents peuvent être hétérogènes et sont définis par une ou plusieurs variables d'état qui sont initialisées au début des simulations. De nombreux agents du même genre peuvent être décrits par une matrice $N \times S$, dans laquelle N est le nombre d'agents du même type et S le nombre de variables d'état utilisées pour les décrire. Pour un nombre prédéfini d'étapes de temps (T), les agents (tous ou une partie d'entre eux) commencent à interagir les uns avec les autres. Les interactions se produisent à chaque étape de temps (t), et leur résultat est une fonction des variables d'état de tous les agents qui interagissent. À la fin de chaque étape, les agents mettent à jour leurs variables d'état, soit en fonction des interactions subies à l'étape t, soit en fonction de quantités prédéfinies.

La publication scientifique est un système dans lequel les auteurs, les relecteurs et les éditeurs interagissent à travers des articles et des revues scientifiques. Tous les agents sont hétérogènes et interagissent en fonction des décisions qu'ils prennent de façon autonome. En outre, les interactions complexes sur lesquelles est basé le système ne garantissent pas un comportement linéaire de celui-ci. Ainsi, le système de publication scientifique est un système complexe et peut être étudié à l'aide d'une modélisation par agents.

Article 2

Complex systems approach to scientific publication and peer-review system: development of an agent-based model calibrated with empirical data

Dans cet article, je présente le développement d'un modèle par agents du système de publication scientifique et d'évaluation par les pairs que j'ai calibré avec des données venant du domaine biomédical. C'est le modèle de base que j'utiliserai pour comparer l'effet d'interventions et alternatives au système dans un travail présenté ultérieurement. J'ai modélisé les chercheurs, les manuscrits et les revues scientifiques en tant qu'agents. Les chercheurs ont été caractérisés par leur niveau scientifique et leurs ressources, les manuscrits selon leur valeur scientifique (Q score) et leurs revues par leur réputation et leurs seuils d'acceptation et de rejet. La système de publication scientifique et d'évaluation par les pairs ont été modélisés comme dans la réalité. Dans mon modèle, les chercheurs investissent des ressources pour créer des manuscrits et ensuite les soumettent à des revues. Ensuite, les manuscrits peuvent être rejetés par décision éditoriale sans évaluation par les pairs, ou bien peuvent être évalués pas des pairs et ensuite soit resoumis à nouveau dans la même revue après modifications, soit rejetés et soumis à une autre revue.

J'ai recueilli des données telles que le taux d'acceptation, le modèle de resoumission et le nombre total d'articles publiés pour les principales revues du domaine biomédical (105 revues). J'ai calibré le modèle par agents afin que les caractéristiques du système correspondent aux données empiriques. Finalement, j'ai simulé les 105 revues, 25 000 chercheurs et 410 000 manuscrits sur 10 ans, et j'ai évalué des mesures macroscopiques du système en faisant des moyennes sur 20 séries de simulations. Ce modèle par agents peut aider à mieux comprendre les facteurs influençant le système de publication scientifique et d'évaluation par les pairs. Il peut également aider à évaluer et à identifier les systèmes alternatifs d'évaluation par les pairs les plus prometteurs, tels que le partage des commentaires et les évaluations collectives effectuées en ligne.

Article 3

Evaluating alternative systems of peer review: a large-scale agent-based modeling approach to scientific publication

Dans cette étude, j'ai modifié la structure du modèle par agents représentant le système conventionnel d'évaluation par les pairs pour simuler cinq systèmes alternatifs. Premièrement, j'ai considéré un système de publication immédiate dans lequel les articles sont immédiatement disponibles en ligne au moment de la soumission, et les éditeurs considèrent à la fois les commentaires de relecteurs choisis par eux, et les commentaires en ligne de la communauté. Après examen, les manuscrits sont indexés dans les bases de données bibliographiques (acceptation) ou rejetés. Deuxièmement, j'ai considéré un système similaire de publication immédiate mais avec uniquement les commentaires de relecteurs choisis par l'éditeur. Troisièmement, j'ai introduit une légère modification au système de base pour étudier une intervention dans laquelle les manuscrits soumis ne font pas plus d'une série d'évaluations et de resoumissions après modification dans un même journal (*re-review opt-out*). Quatrièmement, j'ai modélisé un système dans lequel un article resoumis dans une nouvelle revue après avoir été rejeté doit inclure les évaluations par les pairs faites antérieurement. Cinquièmement, j'ai modélisé un système dans lequel les manuscrits rejetés sont resoumis avec leurs évaluations antérieures à des revues de facteur d'impact inférieur appartenant à un groupe de revues prédéfini, par exemple partageant le même éditeur (*cascade*).

J'ai utilisé trois différents types de mesures pour comparer ces alternatives au système conventionnel: l'efficacité de l'évaluation par les pairs, l'effort des relecteurs et la diffusion scientifique. L'efficacité de l'évaluation par les pairs correspond au double objectif de l'évaluation par les pairs: d'une part sa capacité à distinguer les manuscrits à publier de ceux qui ne devraient pas être publiés en fonction de leur valeur scientifique (Q score), et d'autre part sa capacité à améliorer le Q score des manuscrits une fois effectuées les modifications demandées par le relecteur. J'ai mesuré l'effort des relecteurs comme le temps total consacré par tous les scientifiques à l'évaluation par les pairs durant une année. Enfin, j'ai mesuré la diffusion scientifique en utilisant le nombre de publications annuelles, les semaines médianes entre la première soumission d'un manuscrit et la décision finale, le Q score moyen de tous les manuscrits publiés, et la publication hebdomadaire moyenne d'information scientifique.

Les deux systèmes de publication immédiate ont publié plus d'information scientifique que le système conventionnel, mais n'ont presque aucun autre avantage. Le re-review opt-out a diminué le temps consacré à l'évaluation par les pairs, mais sa performance en termes de dépistage des papiers de basse valeur scientifique et d'augmentation de la qualité intrinsèque des manuscrits était inférieure à celle du système conventionnel. Les performances des systèmes de partage des révisions étaient supérieures ou égales à celles du système conventionnelle en termes d'efficacité de l'évaluation par les pairs, de diminution de l'effort des relecteurs, et de diffusion scientifique. Ils ont surtout produit une forte diminution du temps total du processus d'évaluation par les pairs. Les analyses de sensibilité ont montré des résultats cohérents à ceux présentés ci-dessus. En conséquence, nous recommandons d'évaluer en priorité des systèmes de partage des évaluations dans des études expérimentales en vie réelle.

Discussion

Dans cette thèse de doctorat, j'ai créé un modèle mathématique pour évaluer l'offre et la demande de l'évaluation des manuscrits scientifiques par les pairs, ainsi que le déséquilibre dans l'effort des chercheurs à conduire ces évaluations. Ensuite, j'ai développé un modèle par agents de l'ensemble du système de publication scientifique et d'évaluation par les pairs, je l'ai calibré avec des données empiriques du domaine biomédical, et je l'ai modifié pour évaluer l'effet de stratégies alternatives d'évaluation par les pairs. J'ai comparé ces système alternatifs avec le système conventionnel pour décider quel alternative est plus bénéfique en termes d'efficacité de l'évaluation par les pairs, d'effort des relecteurs et de diffusion scientifique. Dans tous mes projets, l'approche de modélisation que j'ai suivie s'est concentrée davantage sur les aspects macroscopiques du système. Ils ont été calibrés avec des données empiriques du domaine biomédical afin que leurs résultats reflètent autant que possible la réalité.

Mes résultats contestent la prétention dominante, mais anecdotique, qu'il y aurait une pénurie de relecteurs disponibles pour évaluer des manuscrits en raison de l'augmentation du nombre d'articles publiés. Mon travail montre que, au cours des 26 dernières années, il n'y a jamais eu de pénurie d'offre potentielle de relecteurs par rapport à la demande. En fait, en 2015, l'offre de relecteurs potentiels dépassait la demande de relecteurs nécessités de 15% à 249%. Cependant, il y a toujours eu un déséquilibre important dans la répartition de l'effort entre scientifiques pour l'évaluation par les pairs des manuscrits, avec 20% des scientifiques ayant effectué entre 69% et 94% du travail. J'ai estimé que le temps consacré à l'évaluation par les pairs en 2015 était de 63,4 millions d'heures, dont 18,9 millions d'heures ont été fournies par 5% des relecteurs.

Mon approche de modélisation est la première à combiner les caractéristiques macroscopiques de la publication scientifique avec les caractéristiques microscopiques de l'évaluation par les pairs, tout en étant calibrée avec des données empiriques. La plupart des modèles précédents décrivent l'évaluation par les pairs dans le contexte d'une revue scientifique unique, ce qui ne prend pas en compte l'aspect systémique de celui-ci. En outre, ils étaient basés sur des modèles simplistes du processus, qui, en réalité, est très complexe. Enfin, aucun d'entre eux n'a été calibré avec des données empiriques, donc les résultats des simulations étaient trop abstraits et peut-être très éloignés de la réalité. Dans mon modèle par agents, j'ai essayé de saisir la complexité complète du système et de garder ses résultats aussi proches de la réalité que possible.

Mon premier projet a certaines limites. Tout d'abord, la quantification de l'offre et de la demande potentielles pour l'évaluation par les pairs repose sur certaines hypothèses sur les valeurs de paramètres pour lesquels aucune donnée empirique n'est accessible. Cependant, cela ne concernait qu'un nombre limité de paramètres et j'ai effectué des analyses de sensibilité approfondies pour chacun d'eux. En addition, certaines données ne provenaient pas directement des éditeurs, mais de sources secondaires telles que les publications et les enquêtes. Bien que cela limite mon étude dans une certaine mesure, les données sur le système d'évaluation par les pairs sont très limitées, et toutes les sources utilisées étaient celles de la plus haute qualité disponible au moment où j'ai conduit mon travail. Une autre limitation de cette étude était que je n'ai pas considéré les sous-domaines de la biomédecine ni les interactions individuelles entre les auteurs, les relecteurs et les éditeurs. La raison est que mon objectif était d'étudier quantitativement l'offre et la demande globales dans le système. Les sous-domaines et les interactions à la micro-échelle seraient en effet intéressantes d'étudier, mais nécessitent une approche de modélisation différente comme les modèles par agents.

Le modèle par agents est également limité par le manque de données empiriques pour certains paramètres. Puisque j'ai simulé la publication scientifique complète et le processus d'évaluation par les pairs, il existe certaines parties des données impossibles à obtenir, soit parce qu'elles n'existent pas (la distribution réelle du score Q), soit parce qu'elles sont trop difficiles à collecter (le schéma et le volume de resoumission des manuscrits non publiés). Tout biais provenant d'hypothèses dans le modèle serait répliqué de la même manière dans les systèmes alternatifs et, par conséquent, susceptible de s'annuler lorsqu'on compare les alternatives au système conventionnel. En outre, je n'ai pas pu tester toutes les configurations possibles pour modéliser les systèmes alternatifs d'évaluation par les pairs. Cela nécessiterait trop de temps parce que les moyens d'implémenter chaque système alternatif sont nombreux. Ainsi, j'ai choisi l'ensemble des configurations les plus raisonnables pour cette étude.

Conclusion

Dans cette thèse de doctorat j'ai étudié le système de publication scientifique et d'évaluation par les pairs grâce à des modèles mathématiques et des méthodes issues des sciences des systèmes complexes. J'ai montré que, contrairement à des idées anecdotiques existantes, il n'y a jamais eu de pénurie d'offre potentielle de relecteurs par rapport à la demande. Or, 20% des scientifiques ont effectué entre 69% et 94% du total des évaluations par les pairs. J'ai aussi développé un modèle par agents calibré sur des données empiriques pour étudier des alternatives au système conventionnel d'évaluation par les pairs. Basé sur ce modèle, j'ai montré que des systèmes de partage des évaluations devraient être évalués en priorité dans des études expérimentales en vie réelle.

Table of contents

Introduction

Objectives

The objectives of my PhD project were (i) to identify the burden that peer review has been posing to the scientific community, (ii) to simulate the conventional scientific publication and peer-review systems and (iii) to use a simulated framework to compare its efficiency with alternative systems.

The system of scientific publications

Science is humanity's best method of acquiring knowledge that may be trusted to be correct. The system of scientific publication is supposed to be both the gatekeeper and the vector of dissemination of scientific discoveries. On its core, stands the peer-review system as a gold standard for scientific publication. Peer review, in which a scientific communication (article) is evaluated by other researchers (peers) before being published, is used to make sure that irrelevant or badly conducted science does not get published, while helping to improve the quality of those manuscripts deemed (as) publishable (Rennie 2003; Sense About Science 2004). Moreover, the scientific journals and editors are responsible for making sure that all valid scientific knowledge is published and reaches the appropriate target group in, and out of, the scientific community. Thus, the most important actors of the scientific publication system are the researchers, the scientific journals and the editors, who all interact with each other to achieve the common purpose.

Scientific publication has a long history. The first journal exclusively devoted to science was the Philosophical Transactions of the Royal Society and was first issued in 1665 (Spier 2002). At that time, the journal was publishing only papers hand-picked by its editor, Henry Oldenburg, which were not required to pass through any form of peer evaluation. Publishing in scientific journals started taking a form closer to its current one not before mid 19th century (Rennie 2003).

The last few decades have seen the system rapidly increasing in scale. For instance, it was estimated that in 2006 that there were about 1.35 million papers published in about 24,000 journals (Börk 2008). Nine years later, the number of publications almost doubled (2.5 million) and the number of journals increased to 28,000 (Ware and Mabe 2015). Therefore, scientific publication is, currently, a highly complex system with a large number of interactions between multiple heterogeneous actors.

Even though scientific publishing is old, peer review as a system is much younger than most people may believe. For instance, Nature did not introduce a formal peer-review process until 1967 and the Lancet until 1976. Earlier, scientists were very unfamiliar with peer review to the extent that when (the journal) Physical Review sent a paper co-authored by Einstein and Rosen for peer review, Einstein became very upset because they had not authorized the journal to share their research with other scientists prior to publication (Rennie 2003; Csiszar 2016).

Traditionally, scientific journals are accessible to readers, institutions, libraries and to the public through a subscription-based system. However, after the internet started becoming widely adopted in the early years of the past decade this conventional way of publishing has been challenged. The emergence of internet-only journals and online repositories has introduced a new way of publishing which is free for the readers, but not necessarily for the authors or their funding agencies: open access (OA). Open access is an alternative form of publishing, in which either the authors (or their funders) pay a fee to a journal to have their article published on its website and to be freely accessible, if accepted, (gold OA) or they upload it to an online repository (green OA). The gold OA fee is, in the majority of cases, in the form of article-publishing charge or a membership fee. Journals may waive their fees in cases that authors face serious difficulties in paying them. Papers published through green OA do not pass through peer review before being uploaded to a server, however they are usually submitted to scientific journals in which they are subjected to formal peer review. Nowadays, open access dominates scientific publication and even traditional paper-based journals offer gold OA as an option to authors (Dallmeier-Tiessen et al.2010; Suber 2012; Bohannon 2014).

Different scientific domains, in many cases, have adopted differing cultures of peer review and scientific publication. For instance, in physics and mathematics, it is commonplace for papers to be first posted on a Cornell University online server (ArXiv) before being submitted to a journal and undergoing any form of traditional peer review. Thus, many ArXiV papers are shared and discussed by the community in online forums or social media before being published by a journal. This model was also adopted by biology (through bioRxiv), economics (through RePEc) and other fields (PhilSci-Archive, PsyArXiv, ChemRxiv, MedArXiv etc.). However, the habit of discussing papers before formal peer review is still not as widespread in most scientific domains as in physics and mathematics.

There are several metrics to measure the relative impact of scientific journals: the journal impact factor (IF), the 5-year impact factor, the eigenfactor score, the cited half-life etc. The most famous is the journal impact factor, which is the average citations that a journal's papers received over a two-year period (Garfield 2006; Alberts 2013). Depending on the scientific domain, scientists may try to publish in the journals of the highest possible impact factor, according to the perceived importance of their paper, to try and achieve the highest possible amount of readers. The impact factor of journals in which researchers publish their papers as well as the number of citations they receive may be important factors for tenure and rewards. In general, in different domains researchers publish and/or cite more than others, thus the distribution of impact factors might differ between them. In such cases, the percentile of the respective distribution of impact factors inside a domain, which journals belong to, is also used to rank them.

Publishing and the peer-review process

When scientists finish a study, they summarize and present their findings in a report. This report needs to be communicated to the appropriate audience so that it updates the scientific knowledge of all interested parties. Traditionally, the role of dissemination of scientific reports (papers) is assumed by the scientific journals, which are focused on publishing research on either a specialized or general topic. Authors usually submit to journals that may maximize their audience and journals need to select papers that may maximize the interest and size of their audience. Thus, journals implement screening techniques to make sure the content they publish is valid, relevant and interesting to their audience.

Submitted manuscripts first pass through an in-house screening process, during which an editor is assigned to each of them with the responsibility to decide whether a paper is relevant for the journal. In many journals, this decision is taken at a periodical meeting of the editorial board aided by the opinion of the handling editor. This procedure is usually fast, ranging from some hours to a few days. If a paper is rejected, then it may be resubmitted to another journal. If not, then the editor contacts through email other scientists, who are experts or have previously published in the same topic as the submitted manuscript, and asks them to review it. The invitations, which the editors send, usually contain only the abstract of the paper, and based on that, those receiving them should decide whether they want to review the paper or not and notify the editor of their choice.

Candidate reviewers may reject an invitation to review for various reasons such as not having enough time, the paper is not in their field of expertise etc. (Mulligan et al. 2013). Those who accept to review usually do so as volunteers, even though some journals may choose to reward them with some form of discount coupons. Moreover, some journals publish an annual list of their reviewers' names to thank them. Reviewers may also gain credit for their reviews through certain online recognition platforms e.g. Publons (Review rewards 2014; Warne 2016). Peer review usually is single or double blind, meaning that authors do not know the identity of the reviewers while the identity of the authors may either be hidden (double blind) or not from the reviewers (single blind). The rationale for this, is to minimize the possibility of receiving biased reviews, due to the author's seniority, or even retaliation from authors who received critical evaluations of their work towards the reviewers.

The editors, after gathering enough review reports (typically between 1 and 3, sometimes more), take a decision on whether they will reject the paper or ask the authors to make modifications based on the reviewers' comments and resubmit the manuscript for further evaluation. Papers which are not rejected, are re-evaluated and re-revised as many times as necessary until a final decision for rejection or acceptance is made, though more than 2 or 3 review rounds are not frequent. Accepted papers are included in a next issue of the journal, but nowadays they are very often uploaded on its website much earlier than publication in print. Depending on the domain of the paper, the whole procedure may span from a few months to more than one year, whereas a single review report usually requires only a few hours to be completed.

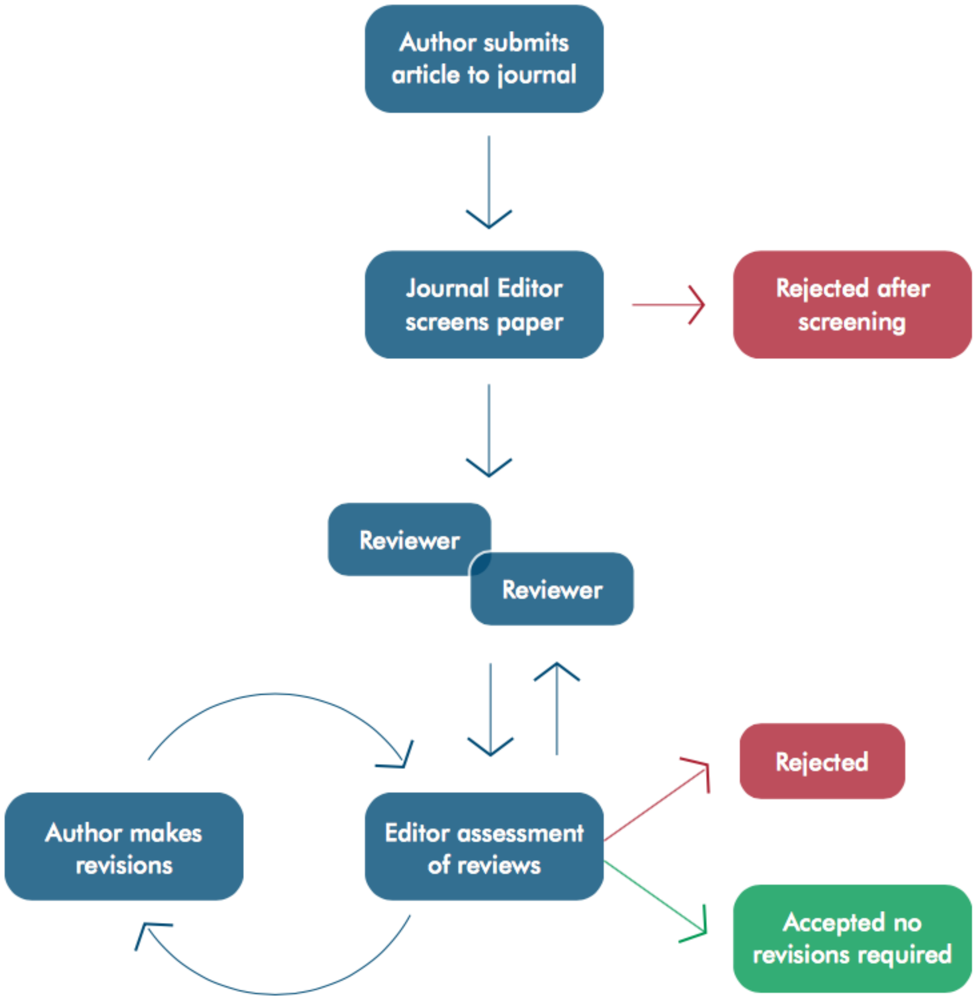


Figure 1: Diagram of the most typical way peer review is currently being held. One can see how authors, journal editors and reviewers interact with each other in the peer-review system (Sense About Science 2012).

Criticism of the peer-review system

Peer review has been recently debated and criticised (Gura 2002; Smith 2006; B. Alberts et al. 2008; Stahel and Moore 2014; Rennie 2016; Csiszar 2016). The huge increase in scientific manuscripts has increased the demand for peer review and potentially introduced a significant burden to the scientific community with a risk of downgraded quality standards on the review reports. It has been stated that, overall scientists need to devote tens of millions of hours per year to perform peer review in biomedicine alone, from which most of it is potentially redundant (i.e. multiple reviews for papers that have already been reviewed once) (Kovanis et al. 2016). In addition, the ability of the system to detect mistakes has also been challenged. For instance, a randomized controlled trial conducted by the British Medical Journal showed that most reviewers could spot 2 to 3 out of 9 major methodological errors in a paper, even after training (Schroter et al. 2004). Another study showed that psychological journals may reject articles that were already published by them, when they were submitted back to them with slightly altered content (Peters & Ceci 1982). Finally, peer review is estimated to cost billions of dollars to scientific institutions annually due to the amount of time that scientists devote to it instead of their normal research activities. (Look and Sparks 2010).

The peer-review process may also be gamed by scientists with unethical motives. A resent scandal of fabricated review reports led to multiple retractions in many journals (Cat et al. 2014; Callaway 2015; Cohen et al. 2016). The scam was revealed when the editors realized that some authors who proposed well-known scientists as reviewers of their papers, were in fact providing them with fake email addresses leading back to themselves. Then, the authors provided very favourable reviews for their own papers, trying to maximize the probability of acceptance. Moreover, another scandal, involving papers published in scientific conferences, revealed that more than one hundred papers, which were automatically generated by a computer had passed through their peer-review process (van Noorden 2014).

After the adoption of open access by the scientific community, various journals that have been considered to be predatory have appeared (Sorokowski et al. 2017). In 2013, Science's editorial staff conducted an experiment and sent to many open access journals around the world a fake paper (Bohannon 2013). While journals should have rejected the paper on sight, many of them accepted the paper sometimes almost instantly after submission. All journals claimed that the paper passed through peer review but without providing any evidence of it. The journals, then, asked the author to pay the regular open access fee, thus aiming to earn easy money by publishing anything.

Interventions and studies on the peer-review system

Even though the implementation of the peer-review process is standard in science, there have been many other alternatives proposed and implemented to some extent. Some journals have experimented with and applied double-blind or open peer review (Blank 1991; van Rooyen et al. 1999; Pöschl 2012; Hopewell et al. 2014; Pontille & Torny 2014; Bruce et al. 2016). Journals have also experimented with providing review training to new authors, because critically appraising a scientific manuscript is considered to be a separate skill than writing one. (Schroter et al. 2004; Houry et al. 2012). Finally, it has also been proposed that manuscripts may be split into parts, such as reporting of outcomes or statistics, which may be reviewed independently by specialized reviewers. Some journals already require an independent statistical review of submitted manuscripts.

Apart from these micro-scale interventions, macroscopic changes have also been proposed and implemented (Walker and Rocha da Silva 2015). First, a system widely adopted in the domain of physics and mathematics, which is the immediate publication of manuscripts before submission to a journal. As described before, in these fields, when authors write a paper they usually upload it on ArXiv and then they submit it to a journal where they follow the standard peer-review process. Thus, all fellow scientists can read the work, discuss and comment on it before publication by the journal. Second, post-publication peer review has gained a lot of support recently (Herron 2012; Hunter 2012). This alternative system allows other scientists to comment and review papers after they have either been posted online or accepted by a journal. Journals currently implementing this system are f1000Research, eLife, the journal of Atmospheric Chemistry and Physics (ACP) etc. Third, another alternative system is one in which an article, after rejection, is resubmitted to another journal together with the previous journal's reviews (Cals et al. 2013; van Noorden 2013). This system aims to eliminate redundant reviews and speed up the decision process.

In the next chapter, I describe my 1^{st} paper, which is about estimating the burden that peer review has been posing to the scientific community. The third chapter contains the description and results of my agent-based model of the current and alternative peer-review systems (2^{nd} & 3^{rd} paper). Finally, the fourth and fifth chapters contain the general discussion and conclusions, respectively.

Paper 1

The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise

Summary

Scientific production has been rapidly increasing in the last decades. A simple search on Pubmed for journal articles returns about 400,000, 700,000 and 1,200,000 results for 1996, 2006 and 2016, respectively. Therefore, the global needs for peer review by the scientific community have increased proportionally. Many prominent scientists have expressed their concerns that the whole system has been overburdening the scientists and may not be sustainable (Arns 2014; Breuning et al. 2015). These claims seem to gain validity if one considers that articles usually undergo more than one submission and multiple rounds of peer review before being published. However, the overall burden of peer review has never been quantified. Thus, there is no evidence on whether the system is actually unsustainable or not. My objective in this study, was to quantify the overall burden that journal peer review has been posing to the field of biomedicine and to estimate how it has been

distributed among the members of the scientific community.

I estimated the annual demand and potential supply of reviews and reviewers using mathematical modeling. I used data mostly pertaining to the field of biomedicine. I downloaded from Pubmed all the information of articles published between 1990 and 2015. Moreover, I retrieved data for the pattern of resubmissions of papers and the time reviewers spend in peer review from an international survey of Elsevier (Mulligan et al. 2013). I obtained the distribution of number of reviews performed by one reviewer in a year from Publons, which is a website that contains data from more than 70,000 reviewers and 10,000 journals. For variables in which no data were available, such as the number of unpublished papers and desk-rejection rates, assumptions and extensive sensitivity analyses were made. Finally, I combined into an equation the desk-rejection rates, average reviewers per paper, number of submitted articles and the probability that articles go to second round of peer review, to estimate the annual demand for reviews and reviewers from 1990 to 2015.

Since there is no commonly agreed definition of who is qualified to review a paper, I defined scenarios to bound the number of researchers who may act as reviewers each year. One scenario was to consider that all first and last co-authors of at least one publication in a year defined a minimum pool of reviewers for that year. We may consider that first co-authors are qualified to review because they have performed research in a field and last co-authors because they have supervised research. As an upper bound for the reviewer pool we considered the number of all co-authors of at least one publication in a given year. The real size of the pool is likely to lie between these two scenarios.

I found that for 2015, across the range of the scenarios investigated, the potential supply exceeded the demand for reviewers and reviews by 15% to 249%. However, 20% of the researchers performed 69% to 94% of the reviews. Among researchers actually contributing to peer review, 70% dedicated 1% or less of their research work-time to peer review while 5% dedicated 13% or more of it. An estimated 63.4 million hours were devoted to peer review in 2015, of which 18.9 million hours were provided by the top 5% of contributing reviewers. These results support that the system is sustainable in terms of volume but emphasize a considerable imbalance in the distribution of the peer-review effort across the scientific community. Thus, the system is not efficient at equalizing the workload of peer review, and therefore, it may be felt as untenable by those overloaded by peer review.

PLOS ONE

RESEARCH ARTICLE

The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise

Michail Kovanis[1,2]*, **Raphaël Porcher**[1,2,3], **Philippe Ravaud**[1,2,3,4,5], **Ludovic Trinquart**[1,4]

1 INSERM U1153, Paris, France, **2** Université Paris Descartes–Sorbonne Paris cité, Paris, France, **3** Assistance Publique-Hôpitaux de Paris, Hôpital Hôtel-Dieu, Centre d'Epidémiologie Clinique, Paris, France, **4** Cochrane France, Paris, France, **5** Department of Epidemiology, Columbia University Mailman School of Public Health, New York, New York, United States of America

* michail.kovanis@aphp.fr

Citation: Kovanis M, Porcher R, Ravaud P, Trinquart L (2016) The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise. PLoS ONE 11(11): e0166387. doi:10.1371/journal.pone.0166387

Editor: Lutz Bornmann, GERMANY

Received: July 11, 2016

Accepted: October 27, 2016

Published: November 10, 2016

Data Availability Statement: All data and analytical results can be found in the accompanying Excel file: http://www.clinicalepidemio.fr/peerreview_burden/. All code is available on github: https://github.com/kovanostra/global-burden-of-peer-review.

Funding: Michail Kovanis is the recipient of a PhD grant from Sorbonne Paris Cité.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The growth in scientific production may threaten the capacity for the scientific community to handle the ever-increasing demand for peer review of scientific publications. There is little evidence regarding the sustainability of the peer-review system and how the scientific community copes with the burden it poses. We used mathematical modeling to estimate the overall quantitative annual demand for peer review and the supply in biomedical research. The modeling was informed by empirical data from various sources in the biomedical domain, including all articles indexed at MEDLINE. We found that for 2015, across a range of scenarios, the supply exceeded by 15% to 249% the demand for reviewers and reviews. However, 20% of the researchers performed 69% to 94% of the reviews. Among researchers actually contributing to peer review, 70% dedicated 1% or less of their research work-time to peer review while 5% dedicated 13% or more of it. An estimated 63.4 million hours were devoted to peer review in 2015, among which 18.9 million hours were provided by the top 5% contributing reviewers. Our results support that the system is sustainable in terms of volume but emphasizes a considerable imbalance in the distribution of the peer-review effort across the scientific community. Finally, various individual interactions between authors, editors and reviewers may reduce to some extent the number of reviewers who are available to editors at any point.

Introduction

The peer-review process of scientific publications became uncomfortable in the scientific community long ago [1, 2]. More recently, several voices have raised concerns about the sustainability of peer review [3–5]. In fact, the number of scientific journals and published articles has increased consistently by about 3% to 3.5% each year; in 2014 alone, about 28,100 peer-reviewed English-language journals published about 2.5 million articles [6]. In the biomedical field, MEDLINE indexed 1.1 million references from more than 5,000 journals in 2015, as compared to about 400,000 and 637,000 references in 1995 and 2005, respectively. Open access and other online journals are a factor in this growth [7].

If articles undergo peer review, the growth in scientific production inevitably puts an increasing burden on the scientific community itself to address the demand for peer review. The process frequently requires second rounds of reviews for a given submission and additional reviews when a manuscript is resubmitted after being rejected. Reviewers typically spend 4 to 5 hours reviewing a paper [8, 9]. The yearly expenditure of peer review is about 2.7 billion US dollars globally [10, 11]. This volume issue may overburden the ability of the scientific community to cope with peer-review duties [5, 12]. However, to our knowledge, we lack concrete evidence about the global demand for reviewers and whether the community self-regulates to cover the demand.

Here we assessed the sustainability of the peer-review system of the scientific publication system in the biomedical domain and how the scientific community is actually coping with the volume of submitted manuscripts.

Methods

Methods summary

We used a mathematical modeling approach, informed by empirical data in the biomedical domain, to compare the quantitative peer-review demand and supply.

We estimated the annual demand as the number of reviews and reviewers required to produce the observed annual number of published articles. The numbers of published articles were derived from MEDLINE for 1990 to 2015 (Fig 1A). We then estimated the corresponding total number of submissions. In fact, an article may be resubmitted multiple times, thus requiring additional reviews. We used the empirical distribution of the number of times papers are resubmitted from data for the biomedical domain in the 2009 Peer Review Survey, an international survey of 4,037 researchers [8]. Moreover, we assumed that 20% of submissions ultimately remained unpublished. We then estimated the corresponding total number of peer reviews (demand for reviews). Some submissions do not require any review, if they are "desk-rejected" after in-house editorial screening. We assumed that the average proportion of desk-rejected papers was 25%. Otherwise, we considered an average of 2.5 reviewers per peer-review round and that 90% of the peer-reviewed submissions went through a second round of peer review [11]. Finally, we estimated the total number of reviewers (demand for reviewers) by using the empirical distribution of individual contributions to the peer review effort (ie, the proportion of reviewers who reviewed 1, 2, 3 etc. papers in a given year) from data for 2015 in the Publons reviewer recognition platform (Fig 1B).

We estimated the annual peer-review supply as the number of potential reviewers and the number of reviews they could perform. Considering that editors typically invite past authors to be peer reviewers, we assumed that potential reviewers in a given year were researchers who co-authored at least one paper that year (Scenario 1). We also considered more stringent scenarios (in terms of co-author consideration to be a potential reviewer) in which candidate reviewers were the first or last authors of any article during the previous 3 years (Scenario 2); the first, second or last authors for the same year (Scenario 3); and the first or last authors for the same year (Scenario 4). For Scenario 2, we arbitrarily chose a time window of 3 years, which however may reflect changes in the databases that editors use to find reviewers. For each scenario, we estimated the number of potential reviewers (supply for reviewers) by counting the unique author occurrences each year from all journal articles indexed in MEDLINE from 1990 to 2015 (Fig 1C). Finally, we estimated the total number of reviews they could perform (supply for reviews) by using the empirical distribution of individual contributions to the peer-review effort.

We estimated the distribution of the proportion of research work-time devoted to peer review. For each researcher, we estimated the total time spent on peer review by using the empirical distribution of the time taken to perform each review from data for the biomedical domain in the 2009 Peer Review Survey (Fig 1D) [8].

Estimation of demand and supply for peer-review

Let us consider N_p the number of articles accepted for publication. Let N_u be the number of articles submitted for publication but that ultimately remain unpublished. We accounted for multiple submissions after rejections, which all occurred within a given year. We assumed that both published and unpublished papers followed the same distribution of resubmissions. Let us define R_i', the proportion of manuscripts submitted exactly i times. The proportion of manuscripts submitted at least i times is $R_i = \sum_{k\geq i} R'_{k-1}$. Then the total number of submissions is:

$$N_s = (N_p + N_u) \times \sum_{i=1}^{I} R_i \times i \quad (1)$$

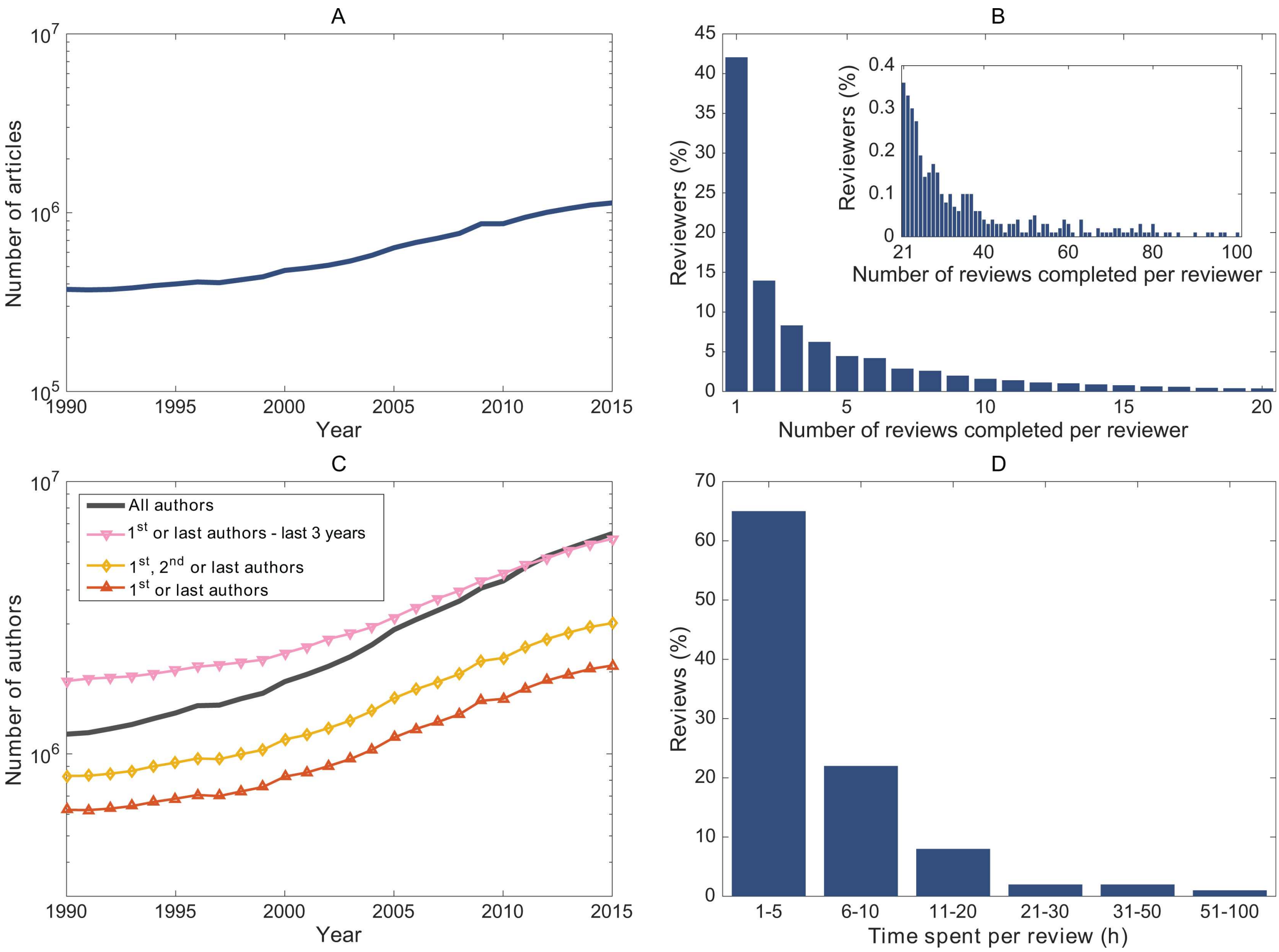


Fig 1. Input distributions and results derived from MEDLINE for peer review in the biomedical domain. (A) Amount of annual publications indexed by MEDLINE and the demand for reviews they generate; (B) Peer-review effort for 2015 provided by Publons. The inset shows the distribution for more than 20 reviews completed per year. Data refer to all scientific domains; (C) Number of authors who published during a given year. Data are from analyzing all annual publications indexed by MEDLINE; (D) Distribution of time spent per review. Data are from Mulligan et al. (2011) and refers to the medical domain.

doi:10.1371/journal.pone.0166387.g001

For simplicity, we set a maximum amount of resubmissions (I). For example, if 5% of papers are submitted once, 10% are submitted twice and 85% are submitted three times, then $R'_1 = 0.05$, $R'_2 = 0.10$, $R'_3 = 0.85$, $R_1 = 1$, $R_2 = 0.95$, and $R_3 = 0.85$. Then, $\sum_{i=1}^{3} \boldsymbol{R_i} \times \boldsymbol{i} = 1 \times 1 + 0.95 \times 2 + 0.85 \times 3 = 5.45$. If we further assume that 800 manuscripts were ultimately published and 200 ultimately unpublished, the total number of submissions is $N_s = 800 \times (1 + 0.95 \times 2 + 0.85 \times 3) + 200 \times (1 + 0.95 \times 2 + 0.85 \times 3) = 1{,}000 \times 5.45 = 5{,}450$ *submissions*.

The distribution of resubmissions of published and unpublished papers might differ, but we can transform it to be the same:

$$N_u{}^0 \times \sum_{i=1}^{I} R_i^0 \times i = N_u{}^0 \times \alpha \times \sum_{i=1}^{I} R_i \times i = N_u \times \sum_{i=1}^{I} R_i \times i \qquad (2)$$

where α is a constant, $\mathrm{N_u}^0 = \frac{N_u}{\alpha}$ the real amount of unpublished papers and $\mathrm{R_i^0}$ the real proportion of papers (re)submitted i times but never published. For example, if $\mathrm{R_1^0} = 1$, $\mathrm{R_2^0} = 0.85$, and $\mathrm{R_3^0} = 0.55$, then $\sum_{i=1}^{3} \mathrm{R_i^0} \times \boldsymbol{i} = 4.35$. If $\mathrm{N_u}^0 = 100$, then the total number of submissions which did not result in a publication is 370. In reality we do not know both $\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{R_i^0} \times \mathrm{i}$ and $\mathrm{N_u}^0$ and it would be impossible to obtain reliable data for them. However, we know $\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{R_i} \times \mathrm{i}$ and we can represent $\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{R_i^0} \times \mathrm{i}$ in terms of it using a constant α. Then, we can group α and $\mathrm{N_u}^0$ into a single constant $\mathrm{N_u}$ and work with Eq 1.

We estimated the annual demand for reviews $\mathrm{N_{reviews}}$ as:

$$N_{reviews} = (1 - d) \times r_s \times \left(N_s + \sum_{i=1}^{I} S_i\right) \qquad (3)$$

where d is the proportion of desk-rejected submissions, r_s the number of reviewers per peer review round and S_i the amount of papers that went to a second round of peer review in their i^{th} (re)submission. We defined S_i as follows:

$$S_i = \beta \times (N_p + N_u) \times R_i \qquad (4)$$

where β is the probability of a second peer-review round per submission that is not desk-rejected.

We can estimate $N_{reviews}$ using a different formula, which this time involves the annual demand for reviewers $N_{reviewers}$.

$$N_{reviews} = N_{reviewers} \times \sum_{j=1}^{J} P_j \times j \qquad (5)$$

where J is the maximum amount of annual reviews that any reviewer performed, j the amount of reviews completed from a reviewer in a given year and P_j the proportion of reviewers who completed j reviews. For example, if 1,000 scientists reviewed at least one paper inside a year, 60% of them performed 1 and 40% of them 2 reviews, then $\mathrm{N_{reviews}} = 1000 \times (0.6 \times 1 + 0.4 \times 2) = 1{,}400$ *reviews*. Since we have two formulas estimating $N_{reviews}$, we can estimate the annual demand for reviewers from their combination:

$$N_{reviewers} = \frac{N_{reviews}}{\sum_{j=1}^{J} P_j \times j} = \frac{(1 - d) \times r_s \times \left(N_s + \sum_{i=1}^{I} S_i\right)}{\sum_{j=1}^{J} P_j \times j} \qquad (6)$$

We defined each researcher's total amount of time available for research as follows:

$$t_{res} = work\ time \times (year - weekends - holidays) \qquad (7)$$

Collection and analysis of data

All data and results can be found in the accompanying Excel file (http://www.clinicalepidemio.fr/peerreview_burden/). We programmed our simulations by using MATLAB (MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Natick, MA, USA). The code is available at https://github.com/kovanostra/global-burden-of-peer-review.

We used data pertaining to the biomedical domain, except to estimate r_s and the distribution of peer-review effort ($\sum_{j=1}^{J} P_j$), for which we used data pertaining to all scientific disciplines. We extracted all records indexed as "journal articles" by MEDLINE from January 1, 1990 to December 31, 2015. We downloaded the xml files for each year separately and parsed them by using a script written in Python (also available on github). We excluded all records with no author name (*e.g.*, less than 0.001% of all articles for 2015) and indexed all authors based on their "LastName", "ForeName" and "Initials". We counted all the unique occurrences of authors by taking into account all these three pieces of information. For missing "ForeName" and/or "Initials", we used only the available fields. We did not use any methods for author name disambiguation for researchers indexed under the same "LastName", "ForeName" and "Initials".[13, 14] We set N_s to be equal to the number of publications for which we identified at least one author.

We assumed that potential reviewers in a given year were researchers who co-authored at least one paper that year (Scenario 1). Then we defined more stringent scenarios (in terms of which co-authors are potential reviewers) whereby candidate reviewers were the first or last authors of any article during the previous 3 years (Scenario 2); the first, second or last authors for the same year (Scenario 3); and the first or last authors for the same year (Scenario 4). For Scenario 2, we arbitrarily chose a time window of 3 years, which however may reflect changes in the databases that editors use to find reviewers. For each scenario, we repeated the same procedure of identifying the unique occurrences of authors as described above. For each scenario, the number of authors obtained was considered to represent the potential supply of reviewers ($N_{reviewers-supply}$) in any given year. We did not account for individual interactions between authors, editors and reviewers which may influence the potential supply of reviewers. We estimated the potential supply of reviews by using the relation $N_{reviews-supply} = N_{reviewers-supply} \times \sum_{j=1}^{J} P_j \times j$.

We obtained $\sum_{i=1}^{I} R_i$ and the empirical distribution of the time taken to perform each review from the 2009 Peer Review Survey, an international survey of 4,037 researchers [8]. Data corresponded to the biomedical domain. We considered r_s to be equal to 2.5 reviewers per peer-review round [11]. We obtained the empirical distribution of individual contributions to the peer-review effort ($\sum_{j=1}^{J} P_j$) for 2015 from the Publons reviewer recognition platform. In Publons, reviewers mainly self-report the reviews they have completed (ie, by forwarding review receipts to them). Publons was launched in 2012 and thus we could not obtain data for all unique years of our analysis. We assumed that the distribution for 2015 was identical for every year from 1990 to 2015.

To our best knowledge, reliable data pertaining to β, N_u and d do not exist. We assumed that 90% of the peer-reviewed submissions went through a second round of peer review ($\beta = 0.9$), the percentage of the finally unpublished papers was equal to the 20% of the total submissions ($N_u = \gamma T_s$, $\gamma = 0.20$) and that the average proportion of papers desk-rejected was 25% ($d = 0.25$). Table A in S1 Appendix presents the values of the previously mentioned parameters.

For each researcher, we estimated the total amount of time available for research t_{res}, taking into account whether the researcher was full or part time. We used empirical data provided by the National Institute of Health and Medical Research of France (INSERM), which pertains to all its researchers. The total time spent in peer review was estimated by sampling the respective empirical distribution over the amount of reviews (j) completed by each reviewer. For example, if 65% of reviews required 1 to 5 hours to complete, 22% of them 6 to 10 etc., then for each review that a reviewer performed we first drew at random the duration range: between 1 and 5 hours with probability 65%, between 6 and 10 with probability 22%, etc. Afterwards, the actual review time was drawn from a uniform distribution over the interval. Comparing the time devoted to peer review with the total time available for research, we derived the proportion of researchers who devoted certain proportions of their time to peer review (full time, 50% or 30% of their annual work-time). For full-time workers, we used *work time* = 8 *hours/day*, *year* = 365 *days* and *weekends* = 104 *days*. We derived the amount of holidays by averaging between 21 OECD countries (*holidays* = 25.3 *days*) [15]. For each full-time employed researcher, we obtained t_{res} = 1,885 *hours* and for part-time researchers t_{res} = 943 *hours* and t_{res} = 566 *hours* for those devoting 50% and 30% of their time to research, respectively.

Sensitivity analyses

We performed 25 sensitivity analyses in addition to our main analysis (Table A in S2 Appendix). We used distributions of peer-review effort other than Publons 2015. Under the same conditions, we obtained the respective distributions from Publons for the years 2013 and 2014. We also used a review effort distribution from only a single journal (Nature Materials 2002–2012). Publons data concerned in total about 70,000 researchers and more than 10,000 journals, whereas data from Nature Materials concerned about 4,500 researchers and a single journal. Finally, we varied the values of the parameters (β, γ, d). We summarized the results of all sensitivity analyses by using the relative difference between the annual number of potential reviewers and the annual demand.

Results

Main analyses

From 1990 to 2015, the demand for reviews and reviewers was always lower than the supply (Fig 2). In 2015, 1.1 million journal articles were indexed by MEDLINE and we estimated that they required about 9.0 million reviews and 1.8 million reviewers. In contrast, depending on the scenario, the annual supply would be between 10 and 30 million reviews and between 2.1 and 6.4 million reviewers. A substantial proportion of researchers do not contribute to the peer-review effort. In fact, the supply exceeded the demand by 249%, 234%, 64% and 15%, depending on the scenario. The peer-review system in its current state seems to absorb the peer-review demand and be sustainable in terms of volume.

If the peer-review effort were split equally among researchers, it would generate a demand for 1.4 to 4.2 yearly reviews per researcher, depending on the scenario. However, we found a considerable imbalance in the peer-review effort in that 20% of researchers perform 69% to 94% of reviews (Fig 3A). The imbalance translates into the time spent on peer review. In all, 70% to 90% of researchers dedicate 1% or less of their research work-time to peer review (Fig 3B). Among researchers actually contributing to peer review, 5% dedicate 13% or more of their research work-time to peer review. In 2015, we estimated that a total of 63.4 million hours were devoted to peer review, among which 18.9 (30%) million hours were provided by the top 5% contributing reviewers.

Sensitivity analyses

When using data from Publons 2014 and 2013, all scenarios to define potential reviewers produced a peer-review supply greater than the demand, except under the most stringent scenario (first or last authors for the same year), in which the demand was higher than the supply before

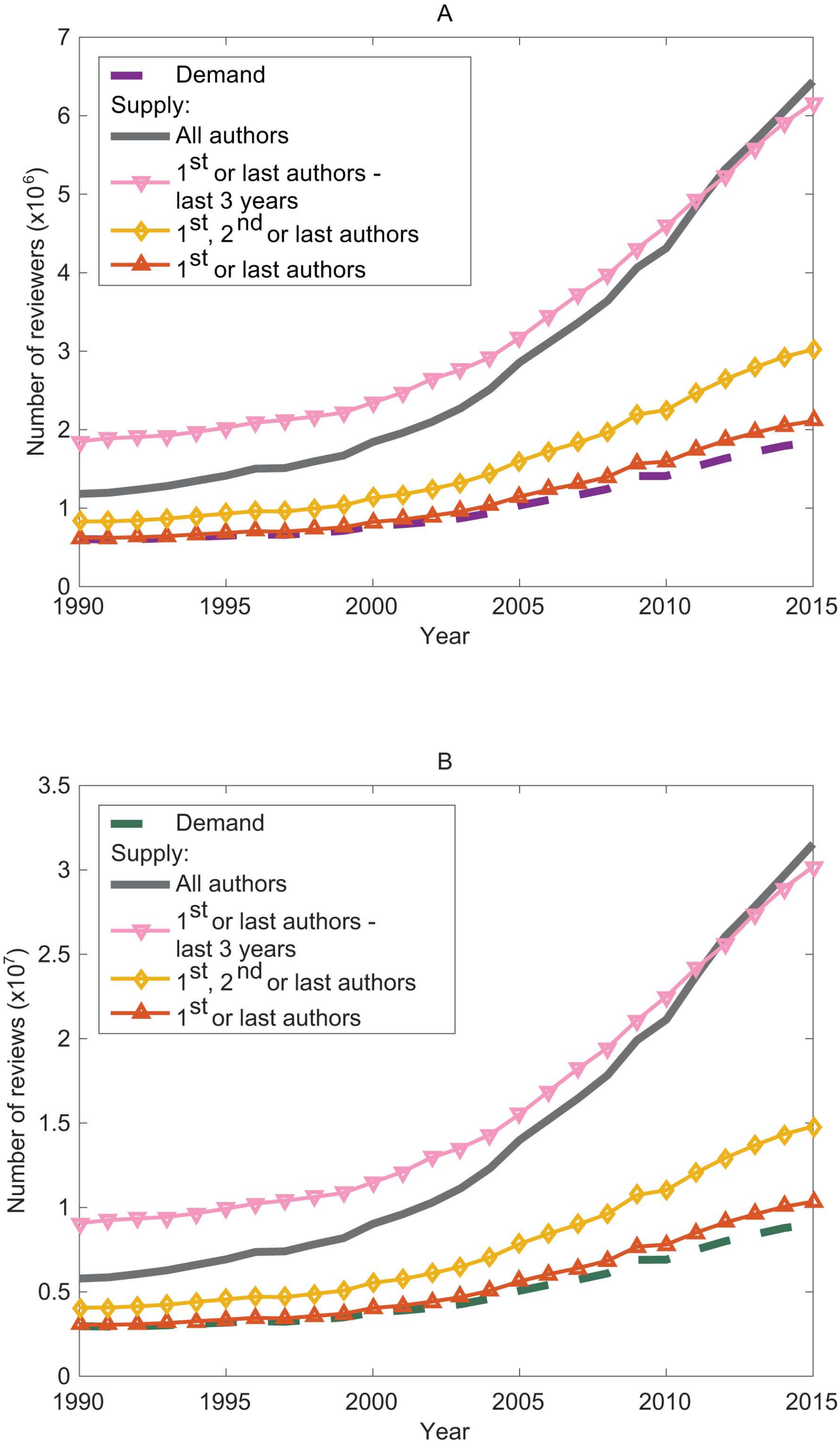


Fig 2. Comparison between supply and demand for reviewers and reviews. (A) Supply and demand for reviewers for all author scenarios. (B) Supply and demand for reviews for all author scenarios.

doi:10.1371/journal.pone.0166387.g002

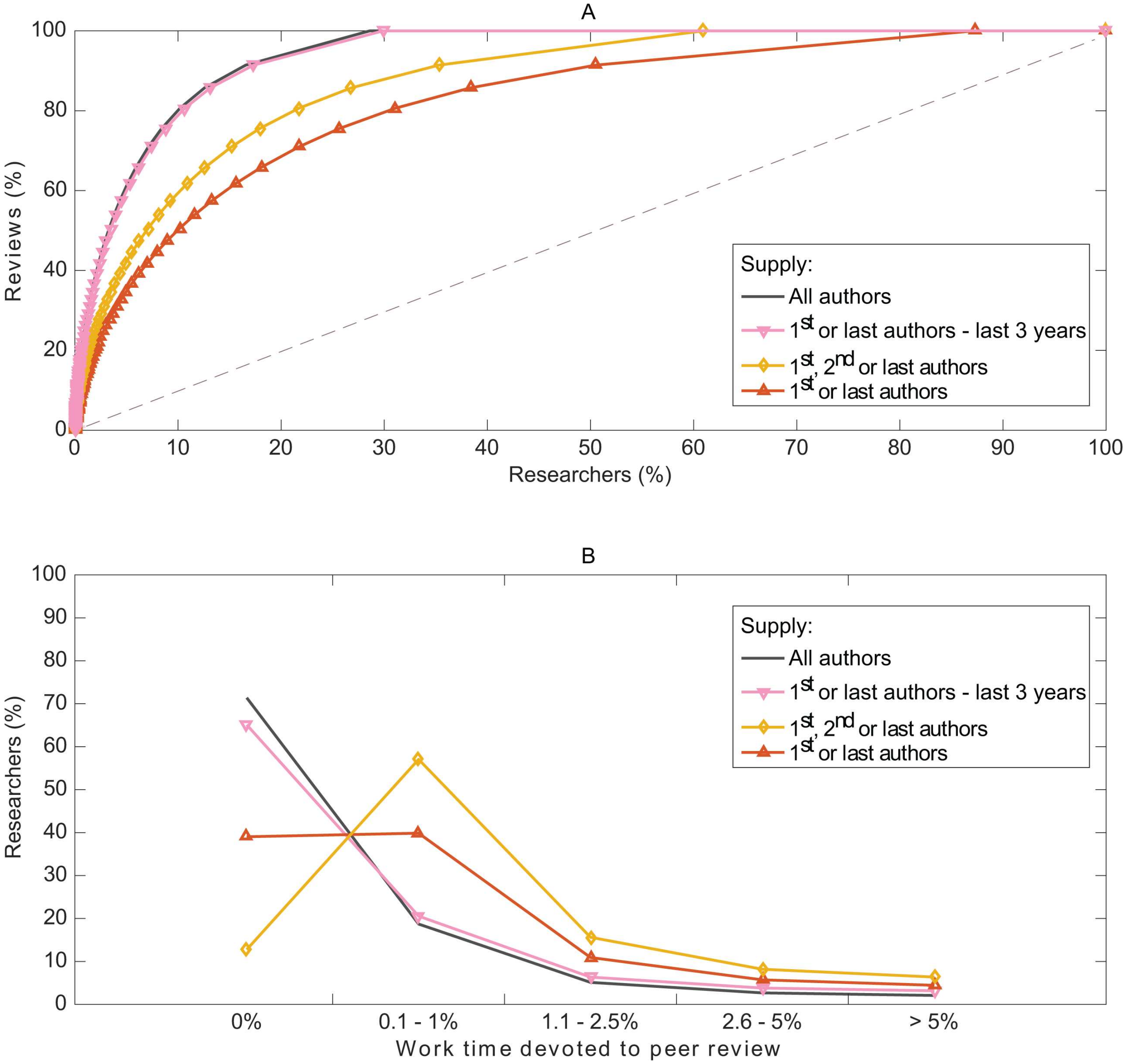


Fig 3. Imbalance in the peer-review effort in terms of workload and work-time. (A) Percentage of authors who complete a certain proportion of the peer-review workload for 2015. (B) Authors' annual percentage of work-time devoted to peer review.

doi:10.1371/journal.pone.0166387.g003

2001 and 2011, respectively (Fig A in S2 Appendix). For 2015 the supply exceeded the demand by 30% and 35%, respectively, when accounting for first, second or last authors and by 0.5% and 5% when accounting for only first or last authors.

When using data from Nature Materials, the scenarios in which all co-authors for the same year and first or last authors for the last 3 years produced a peer-review supply greater than the demand (the second after 1999). As compared with the most stringent scenarios (first, second or last authors and first or last authors for the same year), these data produced a peer-review demand greater than the supply (Fig A in S2 Appendix). For 2015, the supply exceeded the demand by about 30% for both less-stringent scenarios and the demand exceeded the supply by 120% for the most-stringent scenario. However, this is an extreme distribution covering only a single journal.

When varying the values of γ, the peer-review supply was greater than the demand for all scenarios, except for some values > 0.20 when using the most stringent scenario (Figs C and D in S2 Appendix). Variations over the values of β and d also produced a greater supply than demand for all scenarios (except for $d = 0.20$ before 2000) (Figs B and E in S2 Appendix). Almost all sensitivity analyses for the last 3 years produced a surplus in number of available reviewers, even though we compared them to the smallest pool of potential peer reviewers (apart from the one of Nature Materials for the two most stringent scenarios, and for two values of γ in the most stringent scenario). Those that produced a deficit as compared to the most stringent scenario, always produced a surplus as compared to the immediately less-stringent one.

Discussion

Our results challenge recent claims that the growth in published articles may overburden the capacity of the scientific community to absorb the required peer reviews. For the first time, we provide an estimated range for the overall quantitative demand and supply in peer review. The estimation of the quantitative supply we provide refers to the maximum number of reviewers who can be reached by editors according to scenarios and without accounting for individual interactions between authors, editors and reviewers. The scientific community may in fact be able to collectively meet a much higher demand for peer review. This finding is in line with the conclusions of the report of House of Commons Science and Technology Committee and with previous studies in specific journals which showed that peer review was not in crisis [16–19]. However, we showed that a small portion of the scientific community is carrying a disproportionate load of the peer review. These findings are reminiscent of the Pareto principle– 80% of the effects come from 20% of the causes–given that a small number of researchers handles almost all peer reviews. This inequality may be the root of a potentially unmanageable burden. These “peer-review heroes” may be overworked, with risk of downgraded peer-review standards [20].

The geographical distribution of researchers and contributors to the peer-review effort probably explains part of the inequality. In fact, data from two major publishers, Elsevier and Wiley, suggest that, for instance, the proportion of global reviews performed by US researchers is larger than the proportion of global articles they publish. Conversely, Chinese researchers seem to publish twice as many articles as the number they are peer reviewing, despite their willingness to peer review [12, 21].

Peer review should be a collective effort. Reviewing of scientific manuscripts is usually seen as a voluntary and ethical contribution to science, working on a quid pro quo basis. Various reward and incentive systems have been proposed to bolster a more balanced participation in peer-review activities [22, 23]. Reviewer recognition platforms (such as Publons or the Reviewer Recognition Platform) have been launched recently to track and credit peer reviews [24]. Some have suggested offering cash rewards to reviewers or discounts on article processing charges for their future submissions [25]. Such incentives may actually change reviewer motivations and behaviors. Instead, the criteria by which researchers are rewarded for peer-review may be congruent with the more general PQRST system to appraise and reward research, with high-quality transparent peer reviews [26, 27]. Besides, some researchers may be willing to contribute but are never invited. An automated method to improve the matching between submitted articles and the most appropriate candidate peer reviewers may be valuable to the scientific publication system. Such a system could track the number of reviews performed by each author to avoid overburdening them.

Alternative systems of peer review proposed to improve the peer-review system and reduce the burden include “cascade” or “portable” peer review, which would forward the reviews to subsequent journals when papers are resubmitted after being rejected, thus reducing the number of required reviews [28]. Others have suggested re-review opt-out editorial policy or immediate publication with post-publication peer review [29]. A factor that further burdens the peer-review system is the practice of "journal shopping", whereby researchers first target journals with high impact factor and, after rejection, resubmit to journals with gradually lower impact factors. Some initiatives aimed at decreasing journal shopping may contribute significantly to decreasing the overall number of submissions and thus the editorial and peer-review process and the reformatting of manuscripts [30–32].

Here, we focused on journal peer review, but other forms of peer review are likely to impose additional workload on researchers. In particular, the grant peer review system has also been suggested to place a high burden on reviewers. Grant applications may require more work than manuscripts and come in collections at a time because of fixed milestones for submission and deadline systems. Whether the criticism is valid is unclear because empirical evidence concerning the burden on individual researchers and reviewers over time is also scarce [33]. Modeling has been used to address such questions in the grant peer-review system. A recent modeling study from the Office of Extramural Research at the US National Health Institutes (NIH) suggested that the NIH has not tapped the full capacity of the peer-review system [34]. Bollen and colleagues proposed a distributed system and, based on agent-based simulations, showed that the proposed system would result in a similar funding distribution but in less time and cost than the current peer-review system [35, 36].

Our analysis has limitations. First, we assessed the overall quantitative demand and supply and we could not address the qualitative demand and supply. Reviewers are invited by editors on the basis of their expertise in the relevant research area and methodology. “Good” reviewers are likely more solicited for peer review. This situation may explain why the peer-review burden is concentrated on a small portion of researchers. In a survey in political science, 8% of researchers declined requests to review because they considered that they were not sufficiently expert [37]. Moreover, in assessing the supply for peer review, we did not consider that collaborators or scholars from the same institution, for example, may not review each others’ papers or that editors who are also co-authors may not perform additional peer review; these individual interactions between authors, editors and reviewers may reduce to some extent the number of potential reviewers. Instead, we explored this possibility by varying the definition of the pool of potential reviewers according to the ranks of co-authors. Finally, we have not modeled the peer review system as a competitive market economy. In particular, we did not consider the price for peer review and how market forces would apply [38].

Second, we focused on the biomedical literature and our results may not apply to other domains. Even though each discipline has its own characteristics, the biomedical domain accounts for about 44% of the global scientific publications in 2015, and our findings may have implications for domains beyond biomedical research.

Third, we acknowledge that the reliability of our results depends on the data used to inform the modeling. Publons data may not be representative of the true distribution of the peer review effort; registered researchers, who self-report their reviews, may be more intrinsically motivated and more likely to do more reviews than unregistered researchers. To our best knowledge, Publons is the only large-scale source of data about the peer review effort, with data for more than 70,000 reviewers and more than 10,000 journals. We have no data to exclude confidently any selection bias in registered Publons researchers, if any; however, the distribution in Fig 1B shows that 42% of reviewers have reported a single review in 2015. Moreover, Publons has partnered so far with 13 publishers (including Wiley and SAGE), for which registered users automatically receive credit for the reviews they performed (86,910 reviews from 2,676 journals). This finding goes against an overrepresentation of more active reviewers.

We have conducted sensitivity analyses based on data from one specific journal (Nature Materials)[39]. We observed surplus potential reviewer supply when all authors and when the first or last authors across the last 3 years were eligible as reviewers; under more stringent assumptions (1st, 2nd or last and 1st or last author within 1 year), we found a deficit in the reviewer supply. However, researchers are likely to be invited and review for more than one journal; as a consequence, that distribution probably underestimates the effort distribution.

As well, we have used data from the Peer Review Survey to inform the distribution of resubmissions before publication. Although these are the only data about the whole resubmission pattern, they are also limited by self-reporting and a response rate of about 10%. Calcagno et al. previously documented the late submission history of 80,748 articles in biological sciences (self-reporting data with a response rate of 37%) and found that about 75% of published articles were submitted first to the journal that published them.[30].

Another limitation is that our analysis relied on assumptions. However, we restricted these assumptions only to cases when empirical data were, to our best knowledge, not available. In such cases, we set arbitrary but pre-specified values and the values were chosen to reflect realistic scenarios; we performed sensitivity analyses, extensively exploring the parameter space and obtaining results mostly similar to our main analysis (as shown in the S2 Appendix).

One might be interested in analyzing sub-communities of the biomedical system, such as reports of clinical trials. Our search of MEDLINE could have been easily restricted to a smaller selection of articles to reflect these sub-communities. However, summing up the results of all specific sub-communities would give similar results as those obtained from analyzing the whole biomedical domain. Finally, our analysis is also limited by potential issues in the indexation of author names in MEDLINE. Multiple individual researchers can share the same "LastName", "ForeName", "Initials" triplet. Conversely, a given individual researcher could appear as several researchers because of misspellings. We acknowledge that we did not use algorithmic author name disambiguation [13]. The first type of error would lead to underestimating the number of potential reviewers and the second to overestimating the number of potential reviewers. These two types of errors are antagonistic–their effects could be cancelled out–but their impact on our results is difficult to quantify.

In conclusion, the current peer-review system is sustainable in terms of volume but the distribution of the peer-review effort is substantially imbalanced across the scientific community. The evidence base for alternative peer-review systems is still sparse [40, 41]. An evidence-based approach to study peer review, combining computer modeling, experimental studies and sharing of data from journals and publishers, should be encouraged [42–45]. Improvements in peer review will come in response to evidence.[46]

Supporting Information

S1 Appendix. Analytical methods.
(PDF)

S2 Appendix. Sensitivity analyses.
(PDF)

Acknowledgments

We thank INSERM and Publons for providing data; Elise Diard (Cochrane France) for help with formatting figures; and Laura Smales (BioMedEditing, Toronto, ON) for copyediting the manuscript.

Author Contributions

Conceptualization: MK RP PR LT.

Data curation: MK.

Formal analysis: MK.

Funding acquisition: PR.

Investigation: MK LT.

Methodology: MK RP PR LT.

Project administration: PR.

Resources: PR.

Software: MK.

Supervision: RP LT.

Validation: RP PR LT.

Visualization: MK LT.

Writing – original draft: MK LT.

Writing – review & editing: RP PR.

References

1. Lock S. A difficult balance: editorial peer review in medicine. London: Nuffield Provincial Hospitals Trust; 1985.
2. Godlee F, Jefferson T, Callaham M, Clarke J, Altman D, Bastian H, et al. Peer review in health sciences. 2nd ed. London: BMJ books; 2003.
3. Alberts B, Hanson B, Kelner KL. Reviewing peer review. Science. 2008; 321(5885):15. doi: 10.1126/science.1162115 PMID: 18599743.
4. Stahel PF, Moore EE. Peer review for biomedical publications: we can improve the system. BMC Medicine. 2014; 12:179. doi: 10.1186/s12916-014-0179-1 PMID: 25270270
5. Arns M. Open access is tiring out peer reviewers. Nature. 2014; 515(7528):467. doi: 10.1038/515467a PMID: 25428463.
6. Ware M, Mabe M. The STM Report: An overview of scientific and scholarly journal publishing. The Hague, The Netherlands: STM: International Association of Scientific, Technical and Medical Publishers, 2015 March 2015. Report No.
7. Laakso M, Welling P, Bukvova H, Nyman L, Björk B-C, Hedlund T. The Development of Open Access Journal Publishing from 1993 to 2009. PLoS ONE. 2011; 6(6):e20961. doi: 10.1371/journal.pone.0020961 PMID: 21695139
8. Mulligan A, Hall L, Raphael E. Peer review in a changing world: An international study measuring the attitudes of researchers. J Assoc Inf Sci Technol. 2013; 64(1):132–61. doi: 10.1002/asi.22798
9. Ware M. Peer review in scholarly journals: Perspective of the scholarly community-Results from an international study. Inform Serv Use. 2008; 28(2):109–12.
10. Rubriq. How we found 15 million hours of lost time [Blog post] 2013 [updated June 3]. Available from: https://blog.rubriq.com/2013/06/03/how-we-found-15-million-hours-of-lost-time/.

11. Look H, Sparks S. The value of UK HEIs contribution to the publishing process: Summary report: Rightscom Ltd for JISC Collections; 2010. Available from: http://www.jisc-collections.ac.uk/Reports/valueofukhe/.

12. Warne V. Rewarding reviewers–sense or sensibility? A Wiley study explained. Learn Publ. 2016; 29 (1):41–50. doi: 10.1002/leap.1002

13. Schulz C, Mazloumian A, Petersen AM, Penner O, Helbing D. Exploiting citation networks for large-scale author name disambiguation. EPJ Data Science. 2014; 3(1):1–14. doi: 10.1140/epjds/s13688-014-0011-3

14. Lerchenmueller MJ, Sorenson O. Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists. PLoS ONE. 2016; 11(7):e0158731. doi: 10.1371/journal.pone.0158731 PMID: 27367860

15. Ray R, Schmitt J. No-vacation nation USA–a comparison of leave and holiday in OECD countries. EEE Policy Brief. 2007; 3.

16. Vines T, Rieseberg L, Smith H. No crisis in supply of peer reviewers. Nature. 2010; 468(7327):1041. Epub 2010/12/24. doi: 10.1038/4681041a PMID: 21179154.

17. Petchey OL, Fox JW, Haddon L. Imbalance in individual researcher's peer review activities quantified for four British Ecological Society journals, 2003–2010. PLoS ONE. 2014; 9(3):e92896. Epub 2014/03/25. doi: 10.1371/journal.pone.0092896 PONE-D-13-50978 [pii]. PMID: 24658631; PubMed Central PMCID: PMC3962470.

18. Golden M, Schultz DM. Quantifying the Volunteer Effort of Scientific Peer Reviewing. B Am Meteorol Soc. 2012; 93(3):337–45. doi: 10.1175/BAMS-D-11-00129.1

19. Peer review in scientific publications. House of Commons Science and Technology Committee. 2011;1 (par. 134 & 152):46.

20. Summerskill W. In appreciation of the peer-review heroes from 2015. Lancet. 2016; 387(10015):211. doi: 10.1016/s0140-6736(16)00083-0

21. Mulligan A, editor Evolution or Revolution: Changes in Peer Review. STM Frankfurt Conference; 2011 October 11; Frankfurt, Germany.

22. Gasparyan AY, Gerasimov AN, Voronov AA, Kitas GD. Rewarding peer reviewers: maintaining the integrity of science communication. J Korean Med Sci. 2015; 30(4):360–4. Epub 2015/04/02. doi: 10.3346/jkms.2015.30.4.360 PMID: 25829801; PubMed Central PMCID: PMC4366954.

23. Hauser M, Fehr E. An incentive solution to the peer review problem. PLoS Biol. 2007; 5(4):e107. Epub 2007/04/19. 07-PLBI-CC-0206 [pii] doi: 10.1371/journal.pbio.0050107 PMID: 17439298; PubMed Central PMCID: PMC1852148.

24. Review rewards. Nature. 2014; 514(7522):274. doi: 10.1038/514274a PMID: 25318485

25. Chetty R, Saez E, Sándor L. What Policies Increase Prosocial Behavior? An Experiment with Referees at the Journal of Public Economics. J Econ Perspect. 2014; 28(3):169–88. doi: 10.3386/w20290

26. Ioannidis JP, Khoury MJ. Assessing value in biomedical research: the PQRST of appraisal and reward. JAMA. 2014; 312(5):483–4. Epub 2014/06/10. doi: 10.1001/jama.2014.6932 1881107 [pii]. PMID: 24911291; PubMed Central PMCID: PMC4687964.

27. Moher D, Goodman SN, Ioannidis JP. Academic criteria for appointment, promotion and rewards in medical research: where's the evidence? Eur J Clin Invest. 2016; 46(5):383–5. Epub 2016/03/01. doi: 10.1111/eci.12612 PMID: 26924551.

28. Walker R, Rocha da Silva P. Emerging trends in peer review: a survey. Frontiers Neurosci. 2015; 9:169. doi: 10.3389/fnins.2015.00169 PMC4444765. PMID: 26074753

29. Robertson M. Q&A: Re-review opt-out and painless publishing. BMC Biol. 2013; 11:18. Epub 2013/03/01. doi: 10.1186/1741-7007-11-18 1741-7007-11-18 [pii]. PMID: 23445866; PubMed Central PMCID: PMC3584977.

30. Calcagno V, Demoinet E, Gollner K, Guidi L, Ruths D, de Mazancourt C. Flows of Research Manuscripts Among Scientific Journals Reveal Hidden Submission Patterns. Science. 2012; 338 (6110):1065–9. doi: 10.1126/science.1227833 PMID: 23065906

31. Kraus RHS. Peer review: Matchmaker aims to cut journal shopping. Nature. 2016; 531(7595):448–.

32. Powell K. Does it take too long to publish research? Nature. 2016; 530(7589):148. doi: 10.1038/530148a PMID: 26863966

33. Ismail S, Farrands A, Wooding S. Evaluating Grant Peer Review in the Health Sciences: A review of the literature. Santa Monica, CA: RAND Corporation. 2009.

34. Rockey S. Understanding the Capacity of NIH's Peer Review System. NIH Extramural Nexus. 2015.

35. Johan Bollen, David Crandall, Damion Junk, Ying Ding, Katy B. From funding agencies to scientific agency: Collective allocation of science funding as an alternative to peer review. EMBO Rep. 2014; 15:131–3. doi: 10.1002/embr.201338068 PMID: 24397931

36. Bollen Johan, Crandall David, Junk Damion, Ding Ying, Katy B. Collective allocation of science funding: from funding agencies to scientific agency. arXiv:13041067 [physicssoc-ph]. 2013.

37. Breuning M, Backstrom J, Brannon J, Gross B, Widmeier M. Reviewer Fatigue? Why Scholars Decline to Review their Peers' Work. PS: Political Science & Politics. 2015; 48(4):595–600.

38. Ott S, Hebenstreit D. Supply and demand: Apply market forces to peer review. Nature. 2014; 506 (7488):295–. doi: 10.1038/506295b PMID: 24553233

39. A decade in numbers. Nat Mater. 2012; 11(9):743–4.

40. Jefferson T, Wager E, Davidoff F. Measuring the quality of editorial peer review. JAMA. 2002; 287 (21):2786–90. PMID: 12038912

41. Rachel B, Chauvin A, Trinquart L, Ravaud P, Boutron I. Impact of Interventions to Improve the Quality of Peer Review of Biomedical Journals: A Systematic Review and Meta-analysis. Submitted for publication.

42. Kovanis M, Porcher R, Ravaud P, Trinquart L. Complex systems approach to scientific publication and peer-review system: development of an agent-based model calibrated with empirical journal data. Scientometrics. 2016; 106(2):695–715. doi: 10.1007/s11192-015-1800-6 PMID: 26855456

43. Park IU, Peacey MW, Munafo MR. Modelling the effects of subjective and objective decision making in scientific peer review. Nature. 2014; 506(7486):93–6. Epub 2013/12/07. doi: 10.1038/nature12786 PMID: 24305052.

44. Jennings C. Quality and value: the true purpose of peer review. What you can't measure, you can't manage: the need for quantitative indicators in peer review. Nature. 2006. doi: 10.1038/nature05032

45. Siler K, Lee K, Bero L. Measuring the effectiveness of scientific gatekeeping. Proc Natl Acad Sci U S A. 2015; 112(2):360–5. Epub 2014/12/24. doi: 10.1073/pnas.1418218112 1418218112 [pii]. PMID: 25535380; PubMed Central PMCID: PMC4299220.

46. Rennie D. Let's make peer review scientific. Nature. 2016; 535(7610):31–3. doi: 10.1038/535031a PMID: 27383970

Modeling the peer-review system

Introduction

Even though peer review is a cornerstone to scientific publication, its mechanisms have not yet been studied extensively. Although it now seems natural to use peer review, it may not be a system completely based on evidence. As previously stated, only a few trials have been conducted to assess its efficiency as gatekeeper of scientific publication (Rennie 2016; Bruce et al. 2016). In general, ideal randomized controlled trials on scientific publication and peer review may require a large number of researchers evaluating many papers in more than one journals. Extending these trials to evaluate all the alternative systems (or interventions) that have been proposed, may as well complicate their organization.

Instead, one may choose to model the whole scientific publication and peer-review systems and use fast and inexpensive computer simulations to study them, before performing any real-world trials. In the next two papers, I performed computer simulations on the conventional and alternative peer-review systems. Due to the sheer complexity of the systems, I chose to use a method from the domain of complex systems called agent-based modeling. This method allows microscopic modeling of each agent involved and to derive the overall system statistics from the results of their interactions. This approach also permits to perform pre-screenings of alternative interventions and systems and to identify those more likely to improve the conventional system. Therefore, future trials may be guided by the results of the simulations presented here in order to save resources.

Agent-based modeling

Agent-based modeling is a technique in which a system is split and modeled as an ensemble of its composing agents. The agents may be heterogeneous and autonomous, and interact based on a predefined set of rules (Bonabeau 2002). This modeling method is within the complex systems domain. Even though no universal definition of complex systems exists, it is agreed that complex systems are systems, which express emergence and non-linear behavior. This means that the behavior of the system as a whole cannot be expressed as a simple summation of its composing parts (Nicolis and Rouvas-Nicolis 2007). Some very well-known complex systems are the atmosphere, the stock market, the human society and the human brain.

Agent-based modeling is suitable to study complex systems, since it allows modeling of their composing parts one by one. Until now, it has been successfully used to study real life problems in sociology, economy and public health (Epstein 2006; Auchincloss and Diez Roux 2008; Farmer and Foley 2009; Rigotti and Wallace 2015; Chhatwal and He 2015; Marshall and Galea 2015).

To create an agent-based model, one first needs to identify the principal components or agents of the studied system. Agents may be heterogeneous and are defined by one or more state variables, which are initialized at the beginning of the simulations. Many agents of the same kind may be described by a N×S matrix, in which N is the number of agents of the same kind and S the number of state variables used to describe them. The agents (all or a part of them) interact with each other, usually for a predefined number of time steps (T). These interactions happen at each time step (t), and usually, their results come as a function of the state variables of all interacting agents and/or of variables pertaining to their environment. Finally, at the end of each time step the agents update their state variables, either due to rewards or punishments coming from their interactions or by some predefined amounts.

Scientific publication is a system in which authors, reviewers and editors interact through scientific papers and journals. All agents are heterogeneous and interact based on decisions they take autonomously. Also, the behavior of the system is not a simple summation of its parts and not guaranteed to be linear. Therefore, scientific publication is a complex system and it may be studied through agent-based modeling.

Models of peer review

Specific aspects of peer review in scholarly publishing and grant applications have been previously studied by pioneering works using modeling approaches (Martins 2010; Roebber and Schultz 2011; Thurner and Hanel 2011; Allesina 2012; Herron 2012; Squazzoni and Gandelli 2013; Paolucci and Grimaldo 2014; Park et al. 2014; Day 2015).

Squazzoni and Gandelli (2013) modeled a system whereby authors and reviewers interact in the environment of a single journal. They simulated three different scenarios; in the first scenario, the reviewers reciprocated the behavior of previous reviewers towards them; in the second scenario, the reviewers' behavior was not affected by past actions and in the final scenario, the reviewers were reciprocating fair evaluations of their papers. The authors' results suggest that reciprocity can benefit peer review only when inspired by disinterested standards of fairness (Squazzoni and Gandelli 2013).

Thurner and Hanel (2011) used an agent-based model to study the effects of 'correct', 'random', 'rational', 'altruist' and 'misanthropist' behavior of referees on the selection of papers. This model assumed that scientists created papers, which they submitted to two referees who took decisions depending on their behavior. Under a 'correct' behavior, reviewers accepted 'good' and rejected 'bad' papers; under a 'random' behavior, they took decisions randomly; under a 'rational' behavior, they decided based on personal interest; under an 'altruist' behavior, they accepted all papers; and under a 'misanthropist' behavior, they rejected all papers. The authors showed that even a small proportion of 'random' or 'rational' reviewers substantially reduced the average quality of publications (Thurner and Hanel 2011).

Roebber and Schultz (2011) studied peer review in grant applications and more specifically how the rules of the funder and the way decisions are taken affect the scientific community. They created an agent-based model based on Thurner and Hanel (2011) to simulate the funding cycle and assess the efficiency of various proposal-submission strategies. Their results suggest that when the available funding is low then the optimal strategy for the scientists is to submit a high number of grant proposals (Roebber and Schultz 2011).

In another study on peer review of grant proposals, Day (2015) created a model to study how reviewer bias may affect grant-application funding rates. The author simulated a prospective controlled trial in which he introduced bias in different ways to the reviewers' decision-making process. The author found that even a small amount of review bias can lead to statistically significant outcome biases in terms of the number of grant awards (Day 2015).

Paolucci and Grimaldo (2014) redesigned the model of Thurner and Hanel (2011) to replicate their results. In their approach, scientists, conferences and papers interact, whereas reviewers may follow different types of reviewing strategies (correct or rational cheating). The authors show that the obtained results are fragile to small mechanism variations and suggest that exploration at the level of mechanisms is necessary for supporting theoretical statements with simulations (Thurner and Hanel 2011; Paolucci and Grimaldo 2014).

Park et al. (2014) used a model to address problems that arise with reviewers' decisions when their behavior is influenced by the way their peers behave; a phenomenon called herding. To this end, they created three models: one in which subjectivity was introduced in the decisions, another in which all decisions were taken based on objective criteria and lastly one in which all manuscripts were published without any peer review. Their results suggest that there is a high probability for some scientists to submit a paper on topics, which disagree with their initial opinions (herding) and that information transmission is seriously impeded when herding has occurred on a topic. Finally, the chance that more articles would be published on a specific topic increases with the number of papers already published on that topic (Park et al. 2014).

Allesina (2012) modeled a 'classical' setting of the scientific publication system–using 50 journals and 500 researchers–and compared it in terms of efficiency to two alternative settings of the system: 'editorial rejection', in which editors could reject manuscripts after in-house review and 'bidding', in which authors submit their paper to a pool of manuscripts and journals bid for them. The 'editorial rejection' setting raised the publication speed, decreased the burden to the reviewers and provided better control for quality, but raised the rejection rates and the probability of Type I errors as compared to the 'classical' setting. The 'bidding' setting provided faster publication, better distribution of peer-review effort and more publications for authors in 'better' journals, although with higher probability of Type II errors and more burden to the editors than in the 'classical' configuration (Allesina 2012).

Herron (2012) created a model of the traditional peer-review process and compared it with one alternative system; the post-publication peer review. In this model, a group of three expert peer reviewers was compared to reader groups (non-experts) of varying sizes who could evaluate articles. The non-expert reviewers were assumed to be less accurate in their evaluations than the expert reviewers. The simulations showed that even though readers were less accurate than expert reviewers, when the size of the group was higher than 50 persons, the non-expert evaluations were on average more accurate than those of the experts (Herron 2012).

Paper 2

Complex systems approach to scientific publication and peer-review system: development of an agent-based model calibrated with empirical data

Summary

In this paper, I present the development of an agent-based model of the scientific publication and peer-review systems, which I calibrated with empirical journal data in the biomedical and life sciences. This, is going to serve as a base model to which I will compare the effect of alternative systems and interventions in subsequent works. I modeled researchers, research manuscripts and scientific journals as agents. Researchers were characterized by their scientific level and resources, manuscripts by their scientific value (Q score) and journals by their reputation and thresholds of acceptance and rejection. The scientific publication and peer-review procedure was modeled as in reality, with researchers investing resources to create papers and then submitting them to journals. Then, papers were either desk-rejected without external peer review, revised and resubmitted to the same journal or rejected and resubmitted, usually, to another journal.

I collected data for a sample of biomedical and life-sciences journals called the core clinical journals (105 journals) regarding acceptance rates, resubmission patterns and total number of published articles. I fine-tuned the agent-based model so that the respective outputs matched the empirical data. Finally, I simulated the 105 journals, 25,000 researchers and 410,000 manuscripts over 10 years and averaged the results over 20 simulation runs. This agent-based model may help to better understand the determinants of the scientific publication and peer-review systems. It may also help in evaluating the performance of the most promising alternative systems of peer review such as review-sharing and crowdsourcing of online reviews.

Scientometrics
DOI 10.1007/s11192-015-1800-6

Complex systems approach to scientific publication and peer-review system: development of an agent-based model calibrated with empirical journal data

Michail Kovanis[1,2] · **Raphaël Porcher**[1,2,3] · **Philippe Ravaud**[1,2,3,4,5] · **Ludovic Trinquart**[1,4]

Received: 15 July 2015

Abstract Scientific peer-review and publication systems incur a huge burden in terms of costs and time. Innovative alternatives have been proposed to improve the systems, but assessing their impact in experimental studies is not feasible at a systemic level. We developed an agent-based model by adopting a unified view of peer review and publication systems and calibrating it with empirical journal data in the biomedical and life sciences. We modeled researchers, research manuscripts and scientific journals as agents. Researchers were characterized by their scientific level and resources, manuscripts by their scientific value, and journals by their reputation and acceptance or rejection thresholds. These state variables were used in submodels for various processes such as production of articles, submissions to target journals, in-house and external peer review, and resubmissions. We collected data for a sample of biomedical and life sciences journals regarding acceptance rates, resubmission patterns and total number of published articles. We adjusted submodel parameters so that the agent-based model outputs fit these empirical data. We simulated 105 journals, 25,000 researchers and 410,000 manuscripts over 10 years. A mean of 33,600 articles were published per year; 19 % of submitted manuscripts remained unpublished. The mean acceptance rate was 21 % after external peer review and rejection rate 32 % after in-house review; 15 % publications resulted from the first submission, 47 % the second submission and 20 % the third submission. All decisions in the model were mainly driven by the scientific value, whereas journal targeting and persistence in resubmission defined whether a manuscript would be published or abandoned after one or

✉ Michail Kovanis
michail.kovanis@aphp.fr

1 INSERM U1153, 1 Place du Parvis Notre Dame, 75004 Paris, France

2 Université Paris Descartes – Sorbonne Paris Cité, Paris, France

3 Centre d'Epidémiologie Clinique, Hôpital Hôtel-Dieu, Assistance Publique-Hôpitaux de Paris, Paris, France

4 Cochrane France, Paris, France

5 Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA

Published online: 10 December 2015

many rejections. This agent-based model may help in better understanding the determinants of the scientific publication and peer-review systems. It may also help in assessing and identifying the most promising alternative systems of peer review.

Keywords Peer review · Publishing · Computer simulation · Complex systems · Agent-based model

Background and significance

The burden associated with the worldwide scientific production has recently generated much debate and criticism about the sustainability of the established system of scientific publication. The exponential increase in number of manuscripts submitted for publication is much higher than the increase in number of researchers and overburdens the ability of available qualified referees (Ware and Mabe 2015; Gannon 2001; Laakso et al. 2011; Bohannon 2014; Arns 2014; Alberts et al. 2008). Peer-review resources are so scarce that recently the Nature Publishing Group experimented with outsourcing fast-tracked, paid peer review. Moreover, the associated costs are daunting. For example, for the UK higher education institutions alone, peer review would cost more than £110 million annually (Look and Sparks 2010). At the same time, a concern is that the peer-review system may be inefficient at detecting errors and even fraud (Hopewell et al. 2014; Bohannon 2013; Schroter et al. 2008; Stahel and Moore 2014). Most researchers believe that peer review is vital to scientific publication, but it needs to be improved to address all the challenges that arise (Mulligan et al. 2013; Nicholas et al. 2015).

Interventions to improve the system could be targeted to reviewers or the system itself. At the individual level, reviewers could receive special training or authors could be made aware of their identities. Rewarding peer reviewers could provide scientists with incentives to be more involved in peer review activities (Review rewards 2014). Interventions such as cascade peer review (passing reviews of rejected manuscripts to the next editor) or "crowd sourcing" of online reviews (the editor consults online comments along with the reviewers' evaluations) could be implemented at the systemic level (Houry et al. 2012; van Rooyen et al. 1999; Patel 2014; M Ware 2013; Gura 2002; Stahel and Moore 2014; Cals et al. 2013). Assessing the impact of interventions to improve the system would require large-scale experiments, which are complex, costly and sometimes even impossible to perform. In fact, the available randomized controlled trials in the field are few (Rennie and Flanagin 2014).

Scientific publication and peer review need to be studied as a unified system, specifically as a complex system. Computer simulations can reproduce the complete behavior or even uncover data about some elements that are very difficult or impossible to be studied in real life. Agent-based models (ABMs) may be especially useful in this regard.

An ABM aims to simulate and reproduce the behavior and interactions of autonomous real-life agents. The agents interact with each other and their environment, for a complex behavior in the system that differs from the sum of the individual agent behaviors. The characteristics that drive agents' behavior are stored in internal variables and are updated each time some specific conditions are fulfilled or at each time step (Auchincloss and Diez Roux 2008; Galea et al. 2010; Bonabeau 2002; Maglio and Mabry 2011; Epstein 2006). Agent-based modeling is an efficient way to study complex systems (Chhatwal and He

2015; Vespignani 2012; Farmer and Foley 2009; Marshall and Galea 2015; Alberts et al. 2008). It has been successfully used to reproduce and deal with real life problems, especially in public health (Rigotti and Wallace 2015). Previous pioneering studies have used ABM to simulate the peer-review system (F Squazzoni and Gandelli 2013; Paolucci and Grimaldo 2014; Nandi et al. 2013; Lee et al. 2010; Herron 2012; Allesina 2012; Day 2015; Park et al. 2014; Thurner and Hanel 2011).

We aimed to develop an ABM, by adopting a unified view of peer review and publication systems. We attempted to embrace the full complexity of the scientific publication system and use empirical data for calibration. We simulated all the interactions between authors, reviewers and editors and took into account the complete path of scientific manuscripts from submission to the final decision, including resubmissions, rejections after in-house review (without external peer review) and multiple rounds of peer review. We implemented the model in the biomedical and life sciences domain and used empirical data from medical journals for calibration. Our results closely match the real life situation.

Section "Collection of empirical data" of this article describes the sources for our data and section "Modeling the scientific publication and peer-review system", how scientific publication works in real life and the development of our ABM. Section "Calibration procedures and main outputs of the model implementation" describes how we parameterized submodels so that the ABM outputs reproduced the real-life data, and section "Sensitivity analyses" provides the results of our model and sensitivity analyses.

Collection of empirical data

To guide the development and parameterization of the ABM and to perform calibration procedures, we collected empirical data pertaining to the medical domain. We contacted a sample of medical journals to obtain their acceptance rates, with a 40 % positive response rate. We consulted journal websites to obtain the remaining acceptance rates (when available). Finally, we collected the journal impact factors from Journal Citation Reports 2013 and the remaining data from a previously published international survey (Mulligan et al. 2013).

Characterization of journals: survey of editors

Among 119 journals indexed in the MEDLINE Core Clinical Journals subset, we surveyed 105. We excluded journals that invited only submissions ($n = 11$) and those no longer active ($n = 2$); a journal's special edition was considered among the regular issues.

We searched the website for each journal for the contact details of the editor-in-chief or editorial office. On December 5, 2014 we sent an email asking for the number of manuscripts submitted to the journal in 2013, number of manuscripts rejected after in-house review (without external peer review) and number of articles published in 2014. We sent 2 reminders on December 12 and January 22, 2015 and closed our survey on February 1, 2015. We masked the data so that journals could not be matched to their acceptance rates.

We had a response rate of 68 and 40 % for the 105 journals finally provided us with data. For journals that did not provide data, we searched their websites for reported acceptance rates and estimated the number of published articles for 2014 from the Journal Citation Reports 2013. Finally, we collected the acceptance rates for 62 journals and rejection rates after in-house review of 45 journals. We obtained the impact factors for each journal from the Journal Citation Reports 2013 and rescaled them for standardization (Table 1).

Characterization of system processes

We used data from the international survey conducted by Mulligan et al. (2013). The authors contacted 40,000 researchers and obtained 4037 responses. We obtained data for the "medicine and allied health and nursing" domain (565 researchers) for time to final decision for a manuscript (Table 2B) and number of articles that researchers had published (Table 2C). Finally, by directly contacting authors we obtained also the data for the number of submissions up to publication (Table 2A).

Modeling the scientific publication and peer-review system

Description of the system

Scientific publication in its most typical form can be described as a complex system in which researchers interact with each other taking the roles of authors, journal editors and reviewers (Fig. 1) (Brown 2004). Researchers conduct research by using many resources (e.g., grants, research facilities or collaborations). They promote their findings and make them available to the scientific community by reporting them in scholarly manuscripts, which they submit to scientific journals for publication. Decisions on publication are based on multiple factors including the paper's quality, novelty, importance or controversy.

Journals first perform an in-house review to determine whether they will reject a manuscript immediately (e.g., irrelevant to a journal's scope or below quality standards) or send the manuscript for external peer review. In-house review commonly involves the editor-in-chief and members of the editorial board. For the external peer review, the editor solicits external researchers to review articles. On the basis of the editor's and external peer-reviewers' assessments, the editor decides to accept the paper, ask for revision (acceptance is not guaranteed) or reject the manuscript. Revisions require a second or further round of peer review (Wilson 2012). Rejected manuscripts may be resubmitted to other journals or ultimately be abandoned and remain unpublished. Published articles, depending on their impact on the scientific community, help researchers obtain additional resources.

Table 1 Data from MEDLINE Core Clinical Journals

	Data for 2013
Rescaled impact factor($n = 105$)	0.11 ± 0.14 [0.0–1.0]
Acceptance rate ($n = 62$)	0.22 ± 0.11 [0.43–0.59]
Rejection rate after in-house review ($n = 45$)	0.37 ± 0.22 [0.00–0.81]
No. of submissions ($n = 105$)	173,436
No. of rejections after in-house review ($n = 105$)	52,373
No. of published papers ($n = 105$)	32,729

Data are mean $\pm$ SD [min–max] from a survey of journal websites and the Journal Citation Reports 2013

Table 2 Empirical data characterizing the system processes

Process	
A. No. of submissions until publication	No. of articles (n = 565) (%)
1	15
2	47
3	23
4	12
More than 5	4
B. Time to final decision	No. of articles (n = 504) (%)
≤1 week	1
2–3 weeks	5
1–2 months	19
3–6 months	49
>6 months	25
C. Articles	Researchers (n = 4037) (%)
1–5	14
6–10	13
11–20	18
21–50	26
51–100	18
>100	11

Data from Mulligan et al. (2013) international survey and from personal contact with authors

Moreover, researchers benefit from reviewing scientific manuscripts in terms of knowledge.

Agent-based model

We modeled researchers, manuscripts and journals as agents of the scientific publication system from the interactions of their respective state variables (Fig. 1). The researchers could be both authors and reviewers, but editors and journals were modeled as the same agent. The ABM is organized in submodels. The ABM is organized in submodels. Each of the submodels can be parameterized independently. Some submodels pertain to the submission process, including the creation of manuscripts and the targeting of journals for the first submission. Others pertain to the peer review process, including peer review rounds and resubmissions.

Researchers

We characterized N researchers by two state variables: resources $R(t)$ and scientific level $S(t)$ (Squazzoni and Gandelli 2013). The scientific level was defined as $S(t) = R(t) + S_b(t)$, where t the time step and $S_b(t)$ the sum of all the rewards that a researcher can receive to

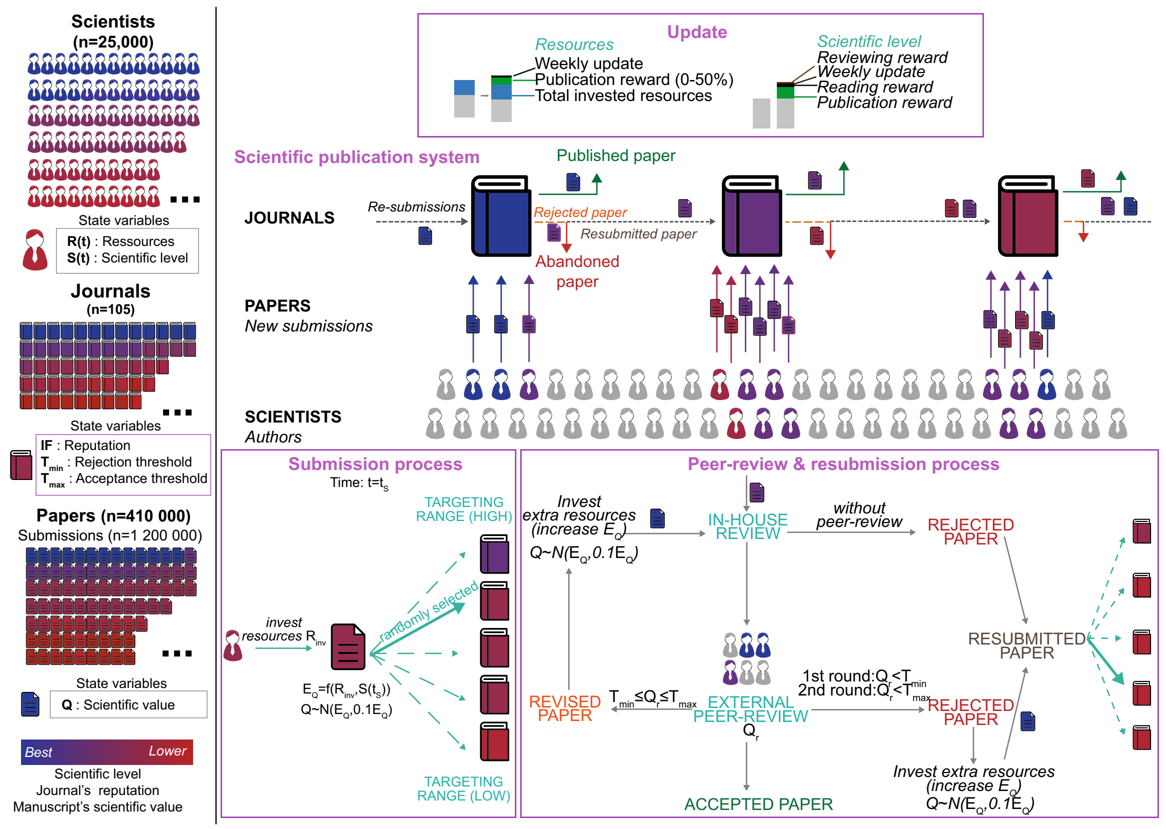


Fig. 1 Description of the agent-based model. The agents and the processes by which our agent-based model operates. Key features are the submodels of the submission and the peer-review process

determine scientific level, as explained at the end of this section. The resources represent all the means that researchers have at their disposal for conducting research. The scientific level expresses a researcher's experience and capacity to conduct better research. In our model, scientific knowledge evolves by a researcher's own research (published articles), the evolution of resources, and from reading and reviewing other manuscripts.

For $t = 0$ we set $S_b(0) = S_p(t)$, where $S_p(t)$ is the cumulative amount of publications per researcher at time t. We initialized $S_p(0)$ (Fig. 2a) using the empirical distribution in Table 2C and set $R(0) = \gamma S_p(0)$ (Fig. 2b) where γ was uniformly distributed over 0.1 and 3. The initial distribution of $S(0)$ can be seen in Fig. 2c.

Manuscripts

Manuscripts were characterized by the state variable Q, which serves as a proxy for their intrinsic scientific value but also their disruptive, innovative, or controversial nature as well as quality of reporting. At each time step, N_s randomly selected researchers submitted their paper (as detailed in the Calibration section). At the time of submission t_s of their paper, authors would lose an amount of resources R_{inv} associated with the conduct of the research reported in that paper—$0.2R(t_s) \leq R_{\text{inv}} \leq 0.7R(t_s)$. However, for researchers with resources, we set $R(t_s) < R_{\text{min}} = 1$ so that they could not submit any work for publication and had to wait until they obtained more resources.

Each paper had an initial expected quality E_Q defined by both the amount of resources the author invested and the author's scientific level at t_s (F Squazzoni and Gandelli 2013)

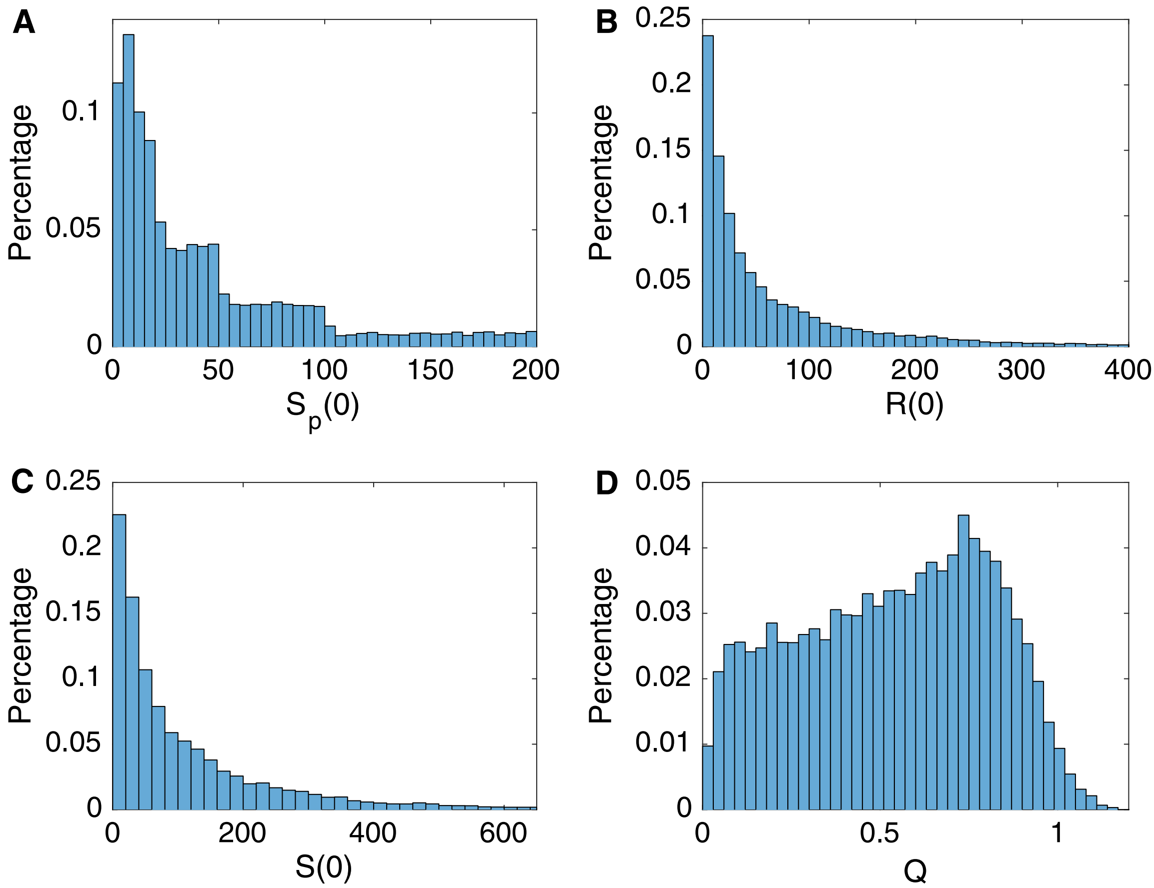


Fig. 2 Distribution of the initial state variables of researchers and articles. Distribution of **a** initial number of published articles per researcher $S_p(t = 0)$, **b** initial amount of resources per researcher $R(t = 0)$, **c** initial scientific level per researcher $S(t = 0)$, and **d** manuscript scientific values (Q scores) when all researchers (N) hypothetically invest half of their available resources at time $t = 0$

$$E_Q = 0.8\frac{0.1R_{\text{inv}}}{0.1R_{\text{inv}} + 1} + 0.2\frac{0.01S(t_s)}{0.01S(t_s) + 1}$$

The Q score was drawn from a normal distribution $Q - N(E_Q, 0.1E_Q)$. This score determines how a researcher chooses a target journal and drives in-house and external peer-review assessments. If all researchers invested half of their initial resources at $t_s = 0$ to create manuscripts, then the distribution of Q scores would be as seen in Fig. 2d.

Journals

We characterized J journals by three state variables: a reputation value [we used rescaled impact factors (Fig. 3a)] and by related rejection or acceptance thresholds, $T^j_{\min} < T^j_{\max}$, $j = 1, \ldots, J$ (Fig. 3b). The reputation and thresholds were used to define how a researcher chose a target journal and if a manuscript was rejected or accepted after in-house or external peer review.

The rejection or acceptance thresholds reflected the ranking of journals by their reputation and were defined by the expected scores of submissions journals receive. For each year, we drew N score values for a fictitious sample of upcoming submissions; we estimated the J-quantiles q^j of this distribution, including the minimum value, and defined $T^j_{\min} = \delta_{\min} q^j + n^j$ and $T^j_{\max} = \delta_{\max} \mathrm{T}^j_{\min} + n^j - C$, where $\delta_{\min}, \delta_{\max}$, and C were constants

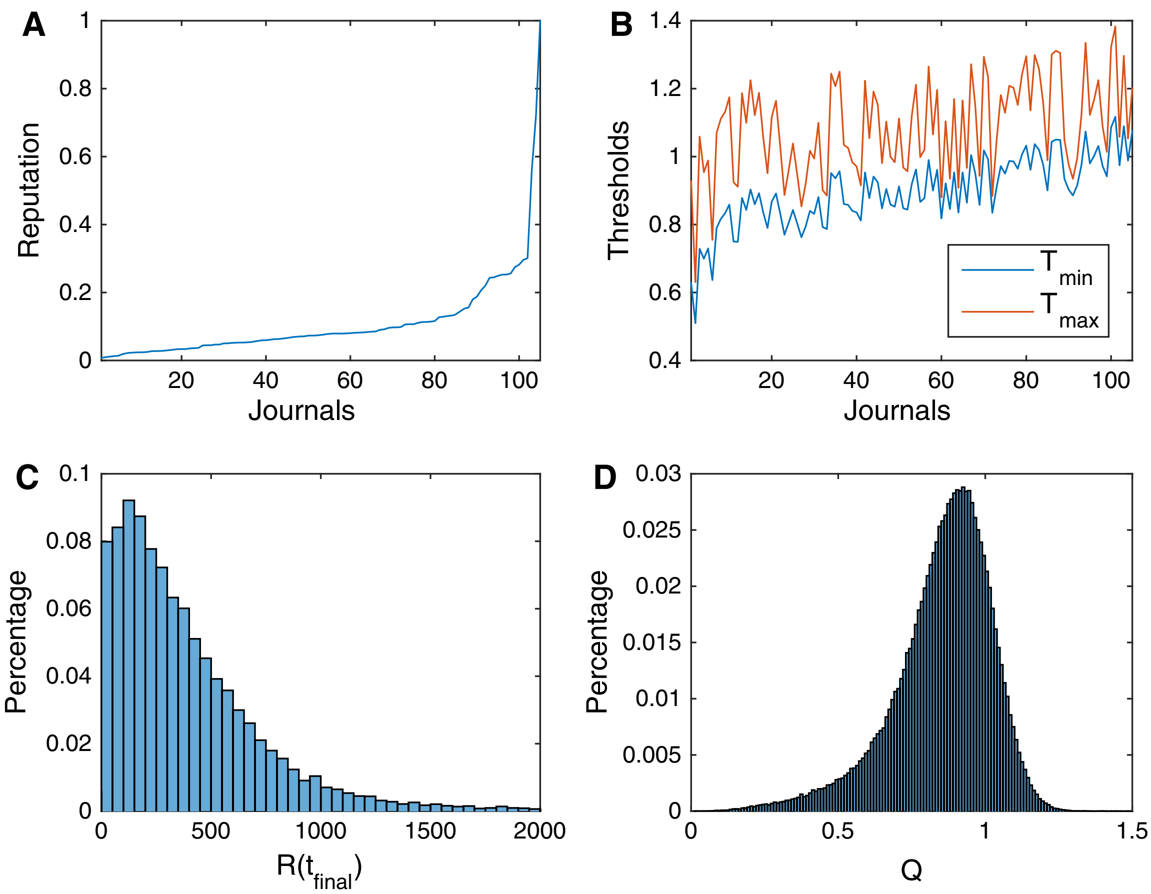


Fig. 3 Distribution of the final state variables of researchers, journals, and manuscripts. Distribution of the **a** journal reputation derived from the rescaled impact factors for 2013, **b** journal rejection and acceptance thresholds $\left(T^j_{\min}, T^j_{\max}, j = 1, \ldots, J\right)$, **c** resources per researcher at the end of the simulations $[R(t = 520)]$, and **d** manuscript final scientific values (Q scores)at the end of simulations

and n^j was random (as detailed in the Calibration section). This definition kept the distribution of acceptance rates insensitive to changes in the distribution of resources.

Journal targeting process

To define how a researcher chose a target journal, we assumed that authors had a general knowledge of journal standards and, given the score, would try to obtain the most recognition from their work. Hence, the journal for the first submission was chosen at random among those with $T^j_{\min}$ within the asymmetrical range $Q - 0.45\varepsilon \le T^j_{\min} \le Q + 0.55\varepsilon$, where $\varepsilon - 2 \times N\left(\frac{Q}{5}, \frac{Q}{20}\right)$. This process resulted in a slight trend of high targeting in every first submission.

In-house and external peer-review process

We drew the editor's assessment of the manuscript Q_e from a uniform distribution over $[0.9Q; 1.1Q]$. If $Q_e < T^j_{\min}$, the manuscript could be rejected without external peer review, depending on the journal's reputation; the likelihood of editorial rejection was larger for journals with larger reputation (as detailed in the Calibration section). If $Q_e \ge T^j_{\min}$, the manuscript was sent for external peer review; two or three reviewers (with 20 %

probability) were randomly selected to their scientific level and the journal's reputation; the top 10 % journals randomly select reviewers among the top 10 % researchers and so on. The reviewers' assessments were defined as $Q_r - N(Q - c, r \times Q)$, where r was a random error and c measured the competitiveness of the reviewer.

The error factor r represents the reliability of the reviewer's assessment. It depended on the amount of time the reviewer spent evaluating the manuscript, the reputation of the journal and the score of the manuscript itself. We assumed that the more time spent on the assessment, the greater the reputation of the journal, and the greater the score, the greater the chance of an accurate assessment. Formally, we defined $r = r_r + r_j - r_Q$, where r_r is the reviewing error, r_j the journal error and r_Q the score error. With 65 % probability, we set $r_t = 0.1$; with 12 %, $r_t = 0.05$; and with 13 %, $r_t = 0.01$ We drew r_j randomly from a uniform distribution over [0; 0.15], where $r_j = 0$ corresponded to the highest reputation journal and $r_j = 0.15$ to the lowest. Finally, $r_Q = 0.05 \times Q$.

The competitiveness factor c depended solely on the reputation of the journal and represents potential reviewer conflict of interest affecting the assessment of the manuscript. We assumed that a competitive behavior would occur more often for journals with higher reputation. The probability of appearance ranged uniformly from 10 to 66 %, where c was drawn randomly from a uniform distribution over [0.01; 0.05].

We randomly selected one of the reviewers' evaluations as a proxy of the editor's opinion. We simulated more than one reviewer to be able to update their scientific levels appropriately. If $Q_r \geq T_{\max}$, the manuscript was accepted and if $Q_r \leq T_{\min}$, it was rejected. When $T_{\min} \leq Q_r < T_{\max}$, the author was asked to revise the manuscript before a second round of peer review.

In the later case, the author invested an extra amount of resources $R_{\mathrm{imp}} - N\left(\frac{8}{60}, \frac{1}{60}\right) \times (R - R_{\mathrm{inv}})$. The cumulative amount of invested resources was used to derive a new Q score as before. The manuscript was re-evaluated by two or three reviewers, randomly selected again, and accepted only if $Q_r \geq T_{\max}$. The Q_r from the second round of peer review was calculated only from the randomly chosen evaluation from the two or three new reviewers.

Following a rejection after in-house review or external peer review, an author could resubmit the manuscript.

Resubmission process

The probability of resubmission P_{res} after a rejection decreased with increasing number of resubmissions r increases, $P_{\mathrm{res}} = P_0^{r-1}$. The P_0 value was defined with the calibration procedure.

If a manuscript was rejected after external peer review, we assumed that the authors could substantially revise it by investing extra resources $R_{\mathrm{imp}} - N\left(\frac{20}{60}, \frac{2}{60}\right) \times \left(R(t_s) - \left(R_{\mathrm{inv}} + \sum_i R^i_{\mathrm{imp}}\right)\right)$, where $R(t_s)$ are the resources before at the time of submission and i the times the author invested extra resources to improve it. If a manuscript was rejected after in-house review, we assumed that authors invested a smaller amount of extra resources $R_{\mathrm{imp}} - N\left(\frac{1}{60}, \frac{0.1}{60}\right) \times \left(R(t_s) - \left(R_{\mathrm{inv}} + \sum_i R^i_{\mathrm{imp}}\right)\right)$.

We assumed that after a first rejection, the authors would target journals of lower reputation than for the first submission. Thus, they randomly selected journals in the (symmetrical this time) range $pQ - 0.5\varepsilon \leq T^j_{\min} \leq pQ + 0.5\varepsilon$, where Q is the initial score of the manuscript and $0 < p < 1$ the targeting of lower reputation journals. This rule allowed for easier acceptance after the second submission, because the score of the manuscript was $>pQ$ after resubmission.

Duration of the peer-review process

For estimating the duration of the peer-review process from submission to final decision, we used the distribution from Table 2B. We assumed that rejection after in-house review occurred within 3 weeks, whereas decisions after one or more rounds of external peer review occurred after ≥ 1 month. When a manuscript is accepted, it takes an extra 1–2 months for publication. Resubmissions occur instantly as the final decision is announced.

Updating of variables

Resources and scientific level were updated at each time step. Resources invested for conducting and reporting research R_{inv} were subtracted at the time of initial submission, whereas the extra resources R_{imp} were subtracted uniformly until the time of a journal's final decision. Thus, a researcher allocated resources to both new research manuscripts and already (re)submitted manuscripts. If the article is published, the author received a reward between 0 and 50 % of the total amount of invested resources, $p \times (R_{\text{inv}} + \sum_i R^i_{\text{imp}}), 0 \leq p \leq 0.5$. If a manuscript remained unpublished, the author would permanently lose all the resources invested.

The scientific level $S(t) = R(t) + S_b(t)$ evolved according to resources and number of published or reviewed manuscripts. In case of publication, the author received a reward for resources in scientific level together with an increase in the number of publications $S_P(t)$. The extra resources invested for revisions were subtracted uniformly from the scientific level until the time of the final decision. The scientific level of a reviewer was credited with a random reward between 0 and 0.001 every time the reviewer completed a review because of knowledge acquired from the paper. Moreover, the scientific level of all researchers was credited with a reward at each time step to reflect the impact of newly published articles, drawn from a normal distribution $N(I, 0.1I)$, where I is the average across all articles published the previous week of $0.1Q_{\text{final}} \times IF^j_{\text{final}}$ (i.e., the quality score of a published article $\times$ the impact factor of the journal that published it). The greater the article quality score and journal impact factor, the higher the chance a researcher would read the article and gain knowledge from it and the larger the reward. Finally, at the end of each week, the researchers received an update to their resources and scientific levels randomly drawn between 0.1 and 1, which reflected an increase of the means to conduct research with time.

Calibration procedures and main outputs of the model implementation

We programmed the model using MATLAB (MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Natick, Massachusetts, United States). The code is available at http://www.clinicalepidemio.fr/peerreview_abm/. We programmed the model with a total population of researchers $N = 25{,}000$ and total population of journals $J = 105$. We ran the simulations for 10 years, with a burn-in period of 1 year for the initialization of the model. Results were averaged over 20 simulations. The main outputs measured were total number of publications per year, proportion of successfully published articles compared to all submissions, proportion of manuscripts revised before being published and proportion of manuscripts for which the peer review process improved their Q score after revision.

We developed our ABM so that its mechanisms resembled those operating in the real-life scientific publication system. We parameterized the model by calibration procedures so that it fit empirically observed data. We considered that this assumption was verified if the model achieved good fit for the distribution of acceptance rates, rejection rates after in-house review, number of submissions until publication, and yearly number of published articles. Goodness-of-fit was assessed by the Anderson–Darling test p values across all runs; we report the minimum and maximum p values.

Distribution of acceptance rates after external peer review

We sorted journals in ascending order by reputation. We generated $n^{(j)} = u^j - z_1^{(j)} + F \times z_2^{(j)}$, with $u^j - U(0.01, 0.20)$, $z_1^j - N(0, 0.45)$ and $z_2^j - N(0, 0.015)$ and $n^{(j)}, z_1^{(j)}$ and $z_2^{(j)}$ order statistics; $F = 1$ for the 20 % highest reputed journals and $F = 0$ for the others. We set $T_{\min}^j = 0.9q^j + n^j$ and $T_{\max}^j = 1.2\mathrm{T}_{\min}^j + (n^j - 0.095)$. We obtained an acceptance rate of 0.21 $\pm$ 0.09, which is almost identical to the one obtained from the survey. Figure 4a shows that the model output fits the empirical distribution of acceptance rates (Anderson–Darling p values [0.63–0.73]).

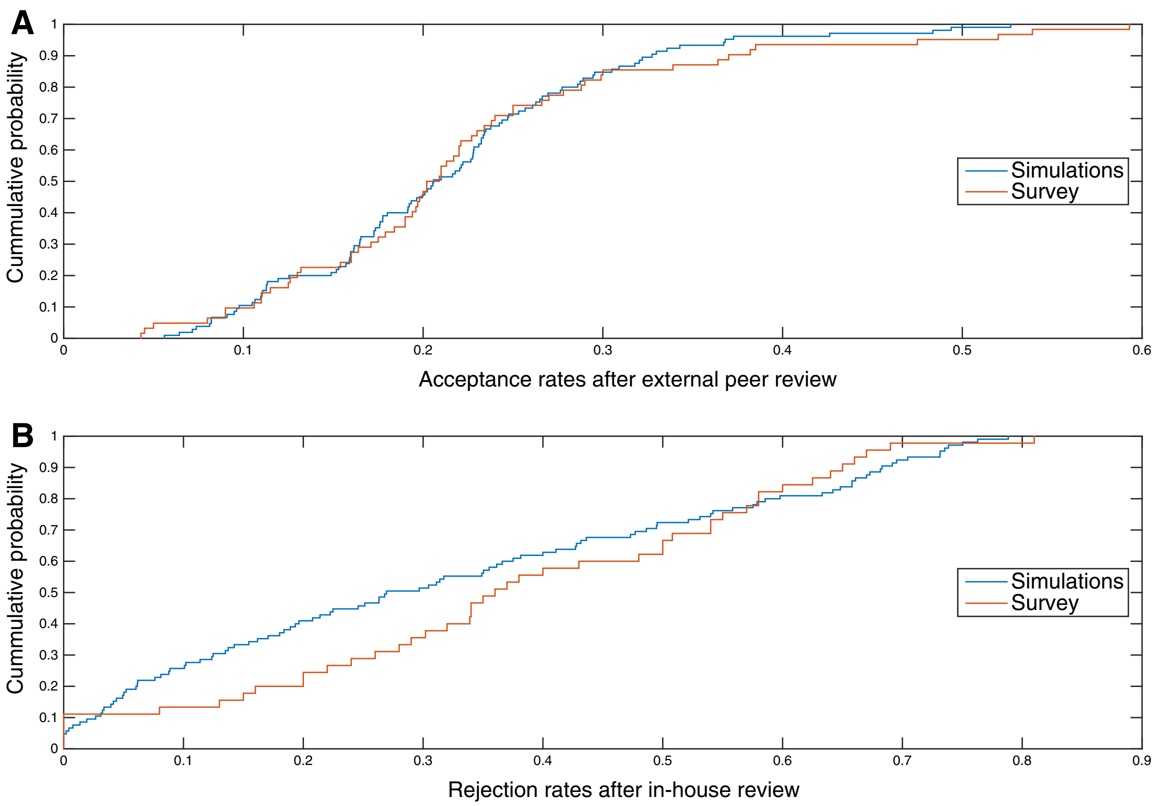


Fig. 4 Calibration of acceptance rates after external peer review and rejection rates after in-house review. Empirical cumulative distribution functions of **a** acceptance rates after external peer review for empirical and simulated data (single run). The model output fits the empirical distribution [Anderson–Darling p values (0.63–0.73) in all runs] and **b** rejection rates after in-house review for empirical and simulated data (single run). The model output fits the empirical distribution [Anderson–Darling p values (0.068, 0.152) in all runs]

Distribution of rejection rates after in-house review

The empirical distribution of rejection rates after in-house review was uniformly spread across impact factors, except for a peak at zero, corresponding to journals that send all submissions for external peer review. To calibrate the distribution, we defined the strictness of the journals as a probability, a linear function of their reputation $\mathrm{Pr}_e^j = \frac{j}{105}$. We also assumed that 20 % of the highest ranked journals would be strict with their editorial policies and would reject everything $< T_{\min}^j$, whereas five of them—excluding the 10 % with the highest reputation—would send everything for external peer review (according to survey data).

Therefore, rejections after in-house review would occur only if $Q_e < T_{\min}^j$ and $\mathrm{Pr} < \mathrm{Pr}_e^j$, where $0 \leq Pr \leq 0.8$ is a random number drawn from a uniform probability distribution. We randomly selected five journals that sent everything for peer review (excluding the top 10 % journals with the highest reputation). This process resulted in uniformly distributed rejection rates after in-house review that match the empirical data as seen in Fig. 4b (mean value [0.32 ± 0.25] and Anderson–Darling p values [0.068, 0.152]).

Number of submissions until publication

To calibrate the distribution of submissions until publication, we set $p = 0.68$, so that authors target journals in the range $0.68Q - 0.5\varepsilon \leq T_{\min}^j \leq 0.68Q + 0.5\varepsilon$ when resubmitting and $P_0 = 0.88$. The results in Table 3 show that the ABM outputs fit the empirical data well.

Total publications per year

For each week, we randomly selected $N_s - N(800, 80)$ authors to invest resources and create manuscripts. The authors produced 33,598 ± 203 manuscripts per year as compared with the 32,729 manuscripts estimated from the empirical data for 2013. From these, 87 % were revised before publication and for 75 % of these, the quality was improved as compared with the empirical values of 92 and 88 %, respectively (Mulligan et al. 2013). Overall, 81 % of the total submissions were finally published, with their mean Q score 0.89 ± 0.13, whereas those unpublished had a mean Q score 0.69 ± 0.20; a relative difference of 29 % (Fig. 5).

Table 3 Comparison of distribution of resubmissions (survey vs agent-base model)

Resubmissions	International survey (%)	Agent-based model (%)
0	14.6	14.89 ± 0.09
1	46.9	47.21 ± 0.22
2	22.6	20.35 ± 0.11
3	11.7	9.41 ± 0.09
4	2.4	4.46 ± 0.05
5	0.5	2.095 ± 0.022
6	0.6	0.94 ± 0.04

Data from Mulligan et al. (2013) international survey and from personal contact with authors

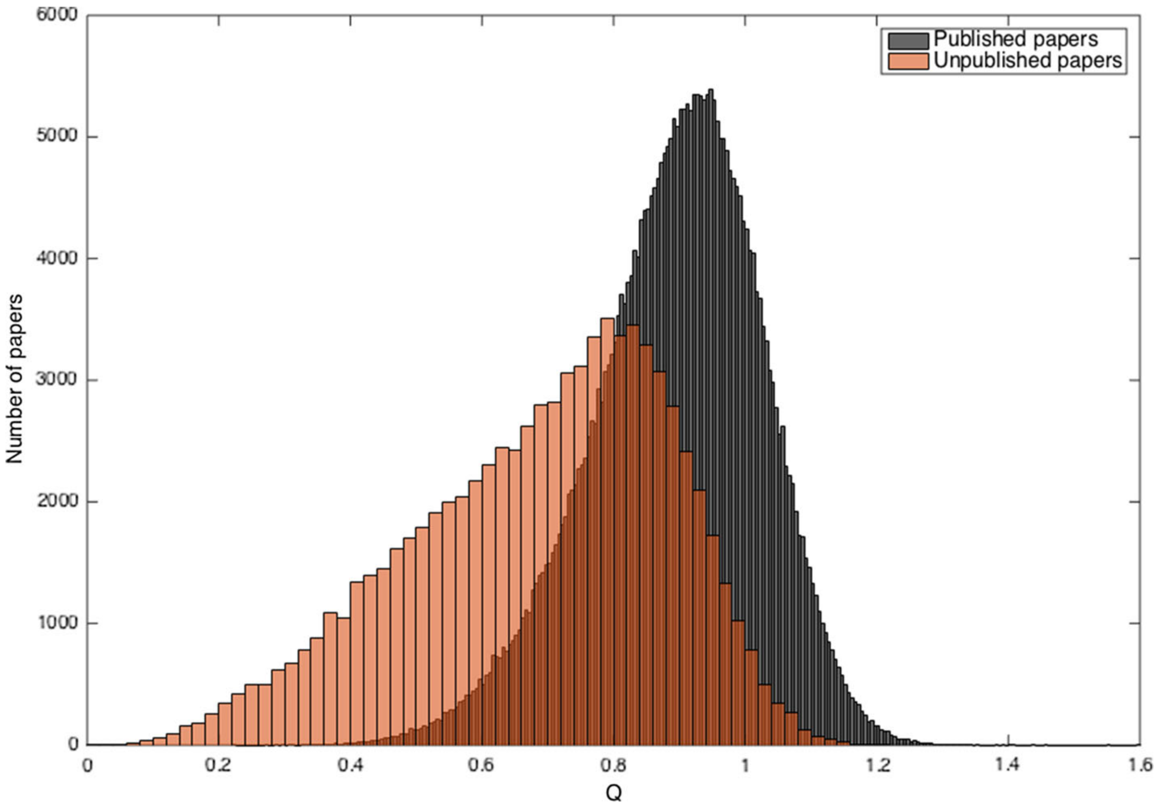


Fig. 5 Distribution of scientific values (*Q* scores) of published and unpublished articles. 81 % of the total submissions were finally published, with their average ***Q*** score 29 % higher from the average ***Q*** score for unpublished manuscripts

Main outputs

Our model fulfills stationarity and ergodicity, and thus, results from a single run do not differ significantly from the average for several runs. We present the results for the main outputs of the model across 20 simulation runs in Table 4.

Sensitivity analyses

We performed two types of sensitivity analyses. First, we selected four variables central to the structure of the model parameters and explored a broad range of values for each so that we could evaluate how they affect the outputs of the ABM. We then performed an extra simulation whereby we evaluated the synergy of the parameter values that maximized difference in the average scores of the published and unpublished manuscripts. Second, we explored various scenarios that incorporate changes in the initial targeting strategy of the authors and in the reviewing behavior of the referees. We compared the results with the standard case for each, to better understand how initial targeting and reviewing strategies can affect the model outputs.

Parameter variation

We performed a sensitivity analysis of the four variables central and varied the targeting when resubmitting (**p**), the volume of weekly submissions ($\mathbf{N_s}$), the probability of

Table 4 Main outputs of the model and their deviations across 20 simulation runs

Yearly publications	Published manuscripts	Revised manuscripts	Improved manuscripts after revisions	Score of published articles	Score of unpublished articles
33,598 ± 203	81.00 ± 0.22	86.69 ± 0.09	75.13 ± 0.11	0.8950 ± 0.0007	0.6880 ± 0.0018
[33,244–34,021]	[80.00–81.00]	[86.57–86.88]	[74.91–75.31]	[0.8940–0.8968]	[0.6855–0.6919]

Data are mean ± SD, in the first row, and [min–max], in the second row, across all simulation runs. We ran the simulations for 10 years, with a burn-in period of 1 year for the initialization of the model, and averaged the results for 20 simulation runs. All runs were calibrated and the outputs varied slightly between the runs resubmission (**P_0**), and the strictness of the in-house reviewing policy (**Pr_{max}**). We measured the impact of the variables on the yearly amount of publications, the proportion of published articles and the difference in the mean scores of published and unpublished manuscripts. For simplicity, we refer to this difference as "score gap".

The scaling of **N_s** linearly increased the value of all the three measured outputs. However, variations in the value of **Pr_{max}** did not have any notable impact on the outputs. The variation of **p** produced the highest difference in the score gap; 40 % relative increase compared to the correct calibration. No parameter variation decreased the difference <10 %. High values of **p** and low values of **P_0** decreased the amount of yearly publications and the percentage of published manuscripts, and vice versa. All results are shown in Table 5.

We performed an extra simulation round for evaluating the extreme scenario, whereby we parameterized the ABM with the values of p and P_0 that produced the maximum score gap. We did not re-parameterize **Pr_{max}**, because variations of its value did not substantially affect the outputs, and **N_s**, because performing simulations for the same amount of scientists submitting per week more than 150 % manuscripts than in the calibrated case would be unrealistic.

The values that affected the score gap the most were $p = 0.5$ and $P_0 = 0.95$ (+40 and +15 % compared to the correct calibration, respectively) and inputted in the model for performing the extra simulation run. The new, more persistent but less ambitious, behavior of the authors when resubmitting resulted in a 55 % increase in the score gap. This increase was produced mainly from the decrease in mean score of the unpublished manuscripts, so with this re-parameterization, the ABM was more capable of low Q score at the screening of papers.

Simulation scenarios

We considered the standard and two additional targeting strategies. In the first strategy, scientists initially submit to journals of lower rejection threshold than they do in the

Table 5 Sensitivity analyses defined by varying four parameters central to structure of the model

Parameter descriptions	Parameter names	Range of variation	Step of variation	Yearly publications	Published manuscripts (%)	Score gap
Targeting when resubmitting	p	[0.9–0.1]	−0.1	[25,911–36,057]	[62–86]	[0.19–0.28]
Volume of weekly submissions	N_s	[400–2000]	200	[16,086–86,131]	[78–83]	[0.18–0.27]
Probability of resubmission	P_0	[0.55–0.95]	0.05	[27,459–35,926]	[66–87]	[0.19–0.23]
Strictness of the in-house reviewing policy	Pr_{max}	[0.1–0.9]	0.1	[33,170–33,770]	[80–81]	[0.20–0.21]

Range of desired outputs [min–max] from sensitivity analysis. The variation of p is presented as a max to min value, because the highest value of p corresponds to the lowest output results and vice versa. Pr_{max} did not substantially affect the outputs, whereas N_s affected them linearly. The variation in p produced the highest score gap (+40 % compared to the correct calibration) standard case ($Q - 0.65\varepsilon \leq T^{j}_{\min} \leq Q + 0.35\varepsilon$). In the second strategy, they target journals of higher rejection threshold ($Q - 0.35\varepsilon \leq T^{j}_{\min} \leq Q + 0.65\varepsilon$). For each of the three strategies, we also considered two additional reviewing scenarios and the standard reviewing scenario. The first scenario assumed that the reviewers would be competitive only if the manuscript they currently review has the same or up to 5 % score as their last published paper. Then they would randomly evaluate its average score as being 5–10 % lower, with all other reviewing errors remaining the same. The second scenario assumed that all evaluations, both of reviewers and editors, are accurate, with no errors.

We compared eight scenarios with the standard model to evaluate how targeting and reviewing affects the system in terms of average number of resubmissions before publication, improvement in papers' scores and increase in the gap in the average scores between published and unpublished papers. No scenario raised the percentage of improved papers after peer review more than 3 %. The fair reviewing strategy increased the score gap the most in all cases (15–16 %), and the rest produced changes varying from −1 to 2 %. Considering the amount of average resubmissions, the competitive case resulted in a decrease ranging from 8 to 15 %, whereas changes from the fair case were insignificant (<5 %). Results in Table 6.

Discussion

Our ABM mimics the properties and functions and addresses different scenarios of behavior and interactions in the scientific publication system. The main strengths of our model are the use of empirical data, which allowed us to produce realistic outputs, and the unified view of evaluation and publishing systems. The main difficulty in the calibration was that we had to reproduce the whole journey of a manuscript from its submission to publication or until the authors give up on submitting it. From empirical data from the biomedical and life sciences domain, we calibrated the model so that the journals do not accept or reject too many manuscripts and so that the manuscripts are not resubmitted more than is required, before being published.

Table 6 Sensitivity analyses defined by the simulation of certain additional scenarios

Initial targeting	Reviewing strategy	Average resubmissions	Improvement after peer review (%)	Average score gap	Relative score gap (%)
Low	Competitive	1.41	74	0.21 [0.69, 0.90]	2
Low	Fair	1.51	78	0.24 [0.67, 0.91]	16
Low	Standard	1.50	75	0.21 [0.69, 0.89]	0
Standard	Competitive	1.46	74	0.21 [0.69, 0.90]	−1
Standard	Fair	1.53	78	0.24 [0.67, 0.91]	15
Standard	Standard	1.56	75	0.21 [0.69, 0.89]	N/A
High	Competitive	1.48	74	0.20 [0.69, 0.90]	−1
High	Fair	1.57	74	0.24 [0.67, 0.91]	16
High	Standard	1.56	75	0.20 [0.69, 0.89]	−1

In this table we see the outputs of the eight scenarios and the relative score gap as compared to the calibrated model [standard–standard]

We obtained an amount of publications very close to that estimated from our survey. This situation allowed us to examine characteristics of manuscripts that remained unpublished and were handled inside the ABM with realistic rules and calibration. In all, 19 % of submissions that received a final decision were never published, and their mean *Q* score significantly differed from that for published articles. A moderate proportion of unpublished manuscripts had *Q* scores close to the high scores of the published articles. This issue is a problem of the scientific publication system, in which editors may sometimes make questionable gatekeeping decisions (Siler et al. 2015). The reasons why an unpublished manuscript considered worthy of publication was not published include poor targeting, mistakes in the in-house or external peer review and lack of persistence in resubmitting the manuscript.

From our sensitivity analyses, variations in the strictness of journals' editorial policies were not able to significantly affect the system. Changes in the amount of weekly submissions linearly affected the model outputs. Behavioral changes in the resubmission strategies of the authors could significantly affect the distribution of *Q* scores of the unpublished manuscripts. The synergetic effect of lower targeted and more persistent resubmissions increased the difference in average scores of published and unpublished papers by 55 %. This finding suggests that the system can publish more papers of higher *Q* score by changes in the resubmission attitudes of authors. However, for producing significant changes in other parts of the system, one needs to consider alternative interventions that will come from structural changes in how the journals and the whole system functions.

From the eight different scenarios, we found that the alternative competitive behavior we introduced reduced the average resubmissions until publication by 8–15 %, without affecting the amount of published articles. The fair scenario produced the highest relative difference in the score gap (15–16 %), which was independent of the initial targeting strategy. However, this difference is still lower than the score gap produced by modifying only the authors' resubmitting behavior. Also, in all cases, the score gap increased or decreased because of the average *Q* score of the unpublished distribution. Finally, the percentage of papers that benefited from peer review did not deviate more than 3 % compared to the standard case for any of these scenarios.

Specific aspects of the peer-review system have previously been studied by pioneering works using ABM approaches (Squazzoni and Gandelli 2013; Park et al. 2014; Allesina 2012; Day 2015; Herron 2012; Paolucci and Grimaldo 2014; Thurner and Hanel 2011). Squazzoni and Gandelli (2013) modeled a system whereby authors and reviewers interact in the environment of a single journal. They simulated three different scenarios; in the first scenario, the reviewers reciprocated the behavior of previous reviewers towards them; in the second scenario, the reviewers' behavior was not affected by past actions and in the final scenario, the reviewers were reciprocating fair evaluations of their papers. The authors' results suggest that reciprocity can benefit peer review only when inspired by disinterested standards of fairness (Squazzoni and Gandelli 2013). Paolucci and Grimaldo (2014) replicated the results of Thurner and Hanel (2011) by using a "redesign" approach. In their approach Scientists, Conferences and Papers interact, whereas reviewers can follow different types of reviewing strategy (Correct or Rational Cheating). The authors show that the obtained results are fragile to small mechanism variations and suggest that exploration at the level of mechanisms is necessary for supporting theoretical statements with simulations (Paolucci and Grimaldo 2014). Allesina (2012) modeled a "classical" setting of the scientific publication system—using 50 journals and 500 researchers—and compared it in terms of efficiency to two alternative settings of the system: "editorial rejection", in which editors could reject manuscripts after in-house review and "bidding", in which authors submit their paper to a pool of manuscripts and journals bid for them. The "editorial rejection" setting raised the publication speed, decreased the burden to the reviewers and provided better control for quality but raised the rejection rates and the probability of Type I errors. The "bidding" setting provided faster publication, better distribution of peer review effort and more publications for authors in better journals although with higher probability of Type II errors and more burden to the editors (Allesina 2012).

However, a holistic approach to evaluate the entire scientific publication system, using empirical data, had not been attempted. Previous studies focused solely on peer review, only a part of our model, or they did not address the full complexity of the system (e.g., large scale of the system, multiple rounds of peer review or revisions of manuscripts after peer review). Despite the continual "risk of brutal oversimplification", we attempted to address the full complexity of the system on a large scale and incorporate empirical data to calibrate its processes (Squazzoni 2010). A reliable base model that better characterizes the standard system must be built and then alternatives to this standard system constructed by comparison because the robustness of inference about the comparison will be influenced by how the standard system is adequately captured by the base model.

The calibration alone was complex, but it was important for describing accurately the base system. We achieved the calibration by "fine-tuning" some microscopic variables to fit empirical data for a limited number of strategically chosen parameters. Alternative systems can be incorporated in the model by making structural changes to some of its submodels. This inclusion will consequently affect the macroscopic outputs. An alternative system would be to crowdsource online reviews and use it along the standard peer review. For implementing this, we need to make additions and modifications to the structure of the submodels of the peer-review process, keeping every other relation and value the same. A structural change could be to allow randomly selected scientists to provide evaluations for a paper, as a form of crowdsourcing of reviews, then the editor to obtain Q_r as the average value of both the regular and the online reviewers comments. However, changes will not be made in the selected values of variables and parameters, only in the relations between them. Since the submodels can be parameterized independently, modifications into any of them do not affect the function of the other. The model will then be able to produce estimates for outputs of systems that have never been implemented in real life. One scenario is how many articles could be published and how fast by an alternative system under the same conditions as the conventional system.

A limitation of our simulations is the use of one-dimensional Q scores. A multidimensional version would treat separately factors such as importance, novelty and controversy arising from the manuscript. For this first exploration, a one-dimensional Q score variable was considered as a satisfactory proxy of all the quality dimensions that a manuscript incorporates. Another limitation is that the peer-review process did not capture the full complexity of interactions as occurs in real life. In next versions of the model, we could increase the complexity of the peer-review process and compare the impact that cooperation and competition between authors, reviewers and editors might have on the system. For example, we could examine in more detail scenarios of conflict of interest and competition for priority between authors and reviewers. We could also make authors spend more resources in the revisions of the paper if the evaluation from the reviewers is closer to the rejection than the acceptance threshold. Furthermore, since reviewers benefit in terms of knowledge from reviewing papers, their rewards could be connected to the Q score of the respective paper. An additional limitation is that our model represents a simplified abstraction of the reality. Arbitrary choices are at some point necessary in order to model real life systems, especially when empirical data are absent. However, to address this limitation, we performed extensive sensitivity analyses, whereby we explored the behavior of the model under several scenarios. A final limitation is that our calibration does not include open-access journals, which can have very different characteristics from traditional style journals. Adding more data, from open-access journals, will increase the accuracy of our calibration measures for scientific publication.

Conclusion

We have developed an ABM that simulates the complexity of scientific publication and peer review and parameterized to fit to certain empirical data coming the biomedical literature. This model produced outputs for both published and unpublished articles. After structural changes to its submodels, we could simulate alternative peer-review systems. The alternative systems that will be produced, depending on the structural changes implemented, will not necessarily be calibrated to the data we used to calibrate the base model. This situation will produce deviations to the measured outputs that will allow us to compare the alternatives to the base system. These comparisons could help highlighting the most promising interventions that may to improve the system and place them under real-life examination.

Acknowledgments Michail Kovanis is the recipient of a PhD grant from Sorbonne Paris Cité. We would like to thank all the editors-in-chief and technical editors who provided us with data for using in calibrating this model. We thank Prof Sandro Galea (Boston University, School of Public Health) for helpful comments on earlier drafts of the manuscript. We thank Elise Diard (Cochrane France) for her help with the complex system diagram and Laura Smales (BioMedEditing, Toronto, ON) for copyediting the manuscript. We would also like to thank the two anonymous reviewers for their useful comments and observations.

References

Alberts, B., Hanson, B., & Kelner, K. L. (2008). Reviewing peer review. *Science, 321*(5885), 15. doi:10.1126/science.1162115.

Allesina, S. (2012). Modeling peer review: An agent-based approach. *Ideas in Ecology and Evolution, 5*(2), 27–35. doi:10.4033/iee.2012.5b.8.f.

Arns, M. (2014). Open access is tiring out peer reviewers. *Nature, 515*(7528), 467. doi:10.1038/515467a.

Auchincloss, A. H., & Diez Roux, A. V. (2008). A new tool for epidemiology: The usefulness of dynamic-agent models in understanding place effects on health. *American Journal of Epidemiology, 168*(1), 1–8. doi:10.1093/aje/kwn118.

Bohannon, J. (2013). Who's afraid of peer review? *Science, 342*(6154), 60–65. doi:10.1126/science.342.6154.60.

Bohannon, J. (2014). Google Scholar wins raves—But can it be trusted? *Science, 343*(6166), 14. doi:10.1126/science.343.6166.14.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America, 99*(Suppl 3), 7280–7287. doi:10.1073/pnas.082080899.

Brown, T. (2004). *Peer review and the acceptance of new scientific ideas*. London: Sense About Science.

Cals, J. W., Mallen, C. D., Glynn, L. G., & Kotz, D. (2013). Should authors submit previous peer-review reports when submitting research papers? Views of general medical journal editors. *The Annals of Family Medicine, 11*(2), 179–181. doi:10.1370/afm.1448.

Chhatwal, J., & He, T. (2015). Economic evaluations with agent-based modelling: An introduction. *PharmacoEconomics, 33*(5), 423–433. doi:10.1007/s40273-015-0254-2.

Day, T. E. (2015). The big consequences of small biases: A simulation of peer review. *Research Policy, 44*(6), 1266–1270. doi:10.1016/j.respol.2015.01.006.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton, NJ: Princeton University Press.

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature, 460*(7256), 685–686. doi:10.1038/460685a.

Galea, S., Riddle, M., & Kaplan, G. A. (2010). Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology, 39*(1), 97–106. doi:10.1093/ije/dyp296.

Gannon, F. (2001). The essential role of peer review. *EMBO Reports, 2*(9), 743. doi:10.1093/embo-reports/kve188.

Gura, T. (2002). Scientific publishing: Peer review, unmasked. *Nature, 416*(6878), 258–260. doi:10.1038/416258a.

Herron, D. M. (2012). Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review. *Surgical Endoscopy, 26*(8), 2275–2280. doi:10.1007/s00464-012-2171-1.

Hopewell, S., Collins, G. S., Boutron, I., Yu, L.-M., Cook, J., Shanyinde, M., et al. (2014). Impact of peer review on reports of randomised trials published in open peer review journals: Retrospective before and after study. *BMJ*. doi:10.1136/bmj.g4145.

Houry, D., Green, S., & Callaham, M. (2012). Does mentoring new peer reviewers improve review quality? A randomized trial. *BMC Medical Education, 12*(1), 83. doi:10.1186/1472-6920-12-83.

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The development of open access journal publishing from 1993 to 2009. *PLoS ONE, 6*(6), e20961. doi:10.1371/journal.pone.0020961.

Lee, B. Y., Brown, S. T., Cooley, P., Grefenstette, J. J., Zimmerman, R. K., Zimmer, S. M., et al. (2010). Vaccination deep into a pandemic wave: Potential mechanisms for a "third wave" and the impact of vaccination. *American Journal of Preventive Medicine, 39*(5), e21–e29. doi:10.1016/j.amepre.2010.07.014.

Look, H., & Sparks, S. (2010). The value of UK HEIs contribution to the publishing process: Summary report. *JISC Collections*. https://www.jisc-collections.ac.uk/Global/report%20on%20HEIs%20noncash%20contribution%20to%20the%20journal%20publishing%20process%20final.pdf .

Maglio, P. P., & Mabry, P. L. (2011). Agent-based models and systems science approaches to public health. *American Journal of Preventive Medicine, 40*(3), 392–394. doi:10.1016/j.amepre.2010.11.010.

Marshall, B. D., & Galea, S. (2015). Formalizing the role of agent-based modeling in causal inference and epidemiology. *American Journal of Epidemiology, 181*(2), 92–99. doi:10.1093/aje/kwu274.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology, 64*(1), 132–161. doi:10.1002/asi.22798.

Nandi, A., Megiddo, I., Prabhakaran, D., & Laxminarayan, R. (2013). An agent-based simulation modelling approach to extended cost-effectiveness analysis of health interventions. *Lancet, 381*, S96. doi:10.1016/S0140-6736(13)61350-1.

Nicholas, D., Watkinson, A., Jamali, H. R., Herman, E., Tenopir, C., Volentine, R., et al. (2015). Peer review: Still king in the digital age. *Learned Publishing, 28*(1), 15–21. doi:10.1087/20150104.

Paolucci, M., & Grimaldo, F. (2014). Mechanism change in a simulation of peer review: From junk support to elitism. *Scientometrics, 99*(3), 663–688. doi:10.1007/s11192-014-1239-1.

Park, I.-U., Peacey, M. W., & Munafo, M. R. (2014). Modelling the effects of subjective and objective decision making in scientific peer review. *Nature, 506*(7486), 93–96. doi:10.1038/nature12786.

Patel, J. (2014). Why training and specialization is needed for peer review: A case study of peer review for randomized controlled trials. *BMC Medicine, 12*, 128. doi:10.1186/s12916-014-0128-z.

Rennie, D., & Flanagin, A. (2014). Research on peer review and biomedical publication: Furthering the quest to improve the quality of reporting. *JAMA, 311*(10), 1019–1020. doi:10.1001/jama.2014.1362.

Review Rewards. (2014). *Nature*. doi:10.1038/514274a. http://www.nature.com/news/review-rewards-1.16138

Rigotti, N. A., & Wallace, R. B. (2015). Using agent-based models to address "wicked problems" like tobacco use: A report from the institute of medicine using agent-based models to address tobacco use. *Annals of Internal Medicine, 163*(6), 469–471. doi:10.7326/m15-1567.

Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. (2008). What errors do peer reviewers detect, and does training improve their ability to detect them? *Journal of the Royal Society of Medicine, 101*(10), 507–514. doi:10.1258/jrsm.2008.080062.

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences of the United States of America, 112*(2), 360–365. doi:10.1073/pnas.1418218112.

Squazzoni, F. (2010). The impact of agent-based models in the social sciences after 15 years of incursions. *History of Economic Ideas, 18*(2), 197–234.

Squazzoni, F., & Gandelli, C. (2013). Opening the black box of peer review: An agent-based model of scientist behaviour. *Journal of Artificial Societies and Social Simulation, 16*(2), 3.

Stahel, P. F., & Moore, E. E. (2014). Peer review for biomedical publications: We can improve the system. *BMC Medicine, 12*, 179. doi:10.1186/s12916-014-0179-1.

Thurner, S., & Hanel, R. (2011). Peer-review in a world with rational scientists: Toward selection of the average. *European Physical Journal B: Condensed Matter and Complex Systems, 84*(4), 707–711. doi:10.1140/epjb/e2011-20545-7.

van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: A randomised trial. *BMJ, 318*(7175), 23–27. doi:10.1136/bmj.318.7175.23.

Vespignani, A. (2012). Modelling dynamical processes in complex socio-technical systems. *Nature Physics, 8*(1), 32–39. doi:10.1038/nphys2160.

Ware, M. (2013). *Peer review: An introduction and guide*. Bristol: Publishing Research Consortium.

Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing* (4th ed.). The Hague: International Association of Scientific, Technical and Medical Publishers.

Wilson, J. (2012). *Peer review: The nuts and bolts*. London: Sense About Science.

Paper 3

Evaluating alternative systems of peer review: a large-scale agent-based modeling approach to scientific publication

Summary

In this study, I modified the structure of the agent-based model for the conventional peer-review system to simulate five alternative systems. First, I considered a system of immediate publication in which articles were immediately available online at the time of submission and editors consider both invited reviews and online comments from the community. After review, manuscripts were either indexed in bibliographical databases (acceptance) or rejected. Second, I considered a similar system of immediate publication but with invited reviews only. Third, I introduced a small modification to the base system to study an intervention in which submitted manuscripts underwent no more than one round of reviews and revisions (re-review opt-out). Fourth, I modeled a system in which rejected papers were resubmitted along with their past reviews to a journal of the authors' choice. Fifth, I modeled a system in which rejected manuscripts were resubmitted

to journals of lower impact factor, within one predefined group of journals (e.g., sharing the same publisher), along with their past reviews.

I used three different types of outcomes to compare all alternatives to the conventional system: the peer-review efficiency, reviewer effort and scientific dissemination. Peer-review efficiency corresponded to the double purpose of peer review. I measured it by the separation of the Q score distributions of the published and unpublished papers and by the mean relative increase of the Q score of all papers after revisions, as compared to that of the first submission. I measured reviewer effort by the total time reviewers devoted to peer review in a year. Finally, I measured scientific dissemination by the number of annual publications, the median weeks between first submission of a paper and the final decision, the average Q score of all papers and the average weekly release of scientific information.

The two systems of immediate publication released more scientific information than the conventional system, but provided almost no other benefit. Re-review opt-out decreased the time reviewers devoted to peer review, but its screening performance of 'poor' papers and the relative increase in papers' intrinsic quality (Q score) due to peer review was lower than in the conventional. The performance of the two review-sharing systems was superior or almost equal to the conventional one in all peer-review efficiency, reviewer effort and scientific dissemination metrics. They importantly decreased the total time of the peer-review process and the total time devoted by reviewers to complete all reviews in a year. As a result, I propose that these two review-sharing systems may be introduced into real-world trials.

Scientometrics
DOI 10.1007/s11192-017-2375-1

Evaluating alternative systems of peer review: a large-scale agent-based modelling approach to scientific publication

Michail Kovanis[1,2] · Ludovic Trinquart[1,4] · Philippe Ravaud[1,2,3,4,5] · Raphaël Porcher[1,2,3,]

Received: 24 November 2016

Abstract The debate on whether the peer-review system is in crisis has been heated recently. A variety of alternative systems have been proposed to improve the system and make it sustainable. However, we lack sufficient evidence and data related to these issues. Here we used a previously developed agent-based model of the scientific publication and peer-review system calibrated with empirical data to compare the efficiency of five alternative peer-review systems with the conventional system. We modelled two systems of immediate publication, with and without online reviews (crowdsourcing), a system with only one round of reviews and revisions allowed (re-review opt-out) and two review-sharing systems in which rejected manuscripts are resubmitted along with their past reviews to any other journal (portable) or to only those of the same publisher but of lower impact factor (cascade). The review-sharing systems outperformed or matched the performance of the conventional one in all peer-review efficiency, reviewer effort and scientific dissemination metrics we used. The systems especially showed a large decrease in total time of the peer-review process and total time devoted by reviewers to complete all reports in a year. The two systems with immediate publication released more scientific information than the conventional one but provided almost no other benefit. Re-review opt-out decreased the time reviewers devoted to peer review but had lower performance on screening papers that should not be published and relative increase in intrinsic quality of papers due to peer review than the conventional system. Sensitivity analyses showed consistent findings to those from our main simulations. We recommend prioritizing a

✉ Michail Kovanis
michail.kovanis@aphp.fr

1 INSERM U1153, 1 place du parvis Notre Dame, 75004 Paris, France

2 Université Paris Descartes – Sorbonne Paris cité, Paris, France

3 Assistance Publique-Hôpitaux de Paris, Hôpital Hôtel-Dieu, Centre d'Epidémiologie Clinique, Paris, France

4 Cochrane France, Paris, France

5 Department of Epidemiology, Columbia University Mailman School of Public Health, New York, USA

Published online: 03 April 2017

system of review-sharing to create a sustainable scientific publication and peer-review system.

Keywords Peer review · Cascade · Portable · Post-publication · Complex systems · Agent-based model

Introduction

The peer-review system is undeniably the gold standard of scientific publication. It serves a double purpose; to screen out bad science and to improve the quality of manuscripts before they are published. However, the scientific community is concerned about the sustainability of the system given the growing number of papers submitted for publication, which puts pressure on the system (Bohannon 2013; Hopewell et al. 2014; Arns 2014; Jennings 2006; Mulligan et al. 2013; Nicholas et al. 2015; Rennie 2016; Sense About Science 2012; Siler et al. 2015; Walker and Rocha da Silva 2015b; Kovanis et al. 2016b).

Much effort has been devoted to proposing alternative systems of peer review or interventions to improve it. However, little effort has focused on testing or evaluating the effectiveness of the alternative systems. Currently, BMC Biology has implemented re-review opt-out, whereby authors are allowed to opt out from a second round of peer review after major revisions to their paper. The journal of Atmospheric Chemistry and Physics has implemented immediate publication upon submission of an article, with online and invited reviews. Philica and F1000 research are also implementing a similar model. Pre-publication servers such as ArXiV or bioRxiv allow researchers to upload their papers before submitting them to a peer-reviewed journal. The Nature and JAMA groups give scientists the option to allow editors of journals within each respective group to discuss rejected manuscripts and to propose submission to another journal of the group (Walker and Rocha da Silva 2015b; Cals et al. 2013; Gura 2002; Houry et al. 2012; Patel 2014; Stahel and Moore 2014; van Rooyen et al. 1999; Ware 2013).

Until 2016, only 22 randomized controlled trials had been conducted to assess peer-review interventions such as double-blind peer review and the addition of a statistical reviewer (Bruce et al. 2016). Studying all the proposed and already-implemented alternatives is not easy. Putting all of them under a real-life test would be costly, time-consuming and sometimes not feasible. Thus, we need approaches such as computer simulations that would allow for quicker screening to identify the most promising alternatives to the peer-review system to be later examined in a real-life test.

Because of the highly complex nature of the scientific publication system, here we used techniques from complex systems modelling, specifically agent-based modelling (ABM), to describe the system. Because of multiple interactions of many heterogeneous and independent agents (e.g., authors, editors, reviewers, papers), this sort of systemic thinking and detailed microscopic modelling was necessary (Galea et al. 2010; Vespignani 2012; Bonabeau 2002; Marshall and Galea 2015). Author, editor and referee behaviour has been extensively studied with ABM and other modelling approaches. Some authors focused on how the number of reviewers, reciprocity, rationality and other motives between referees and authors affect the quality of peer review, and others redesigned models to replicate their results (Bianchi and Squazzoni 2015; Squazzoni and Gandelli 2013; Thurner and Hanel 2011; Paolucci and Grimaldo 2014; Righi and Takács 2017). Others modelled how objectivity and subjectivity in reviewers' decisions macroscopically bias peer review (Park et al. 2014) or estimated the level of bias necessary to affect peer review in grant applications (Day 2015). There have also been attempts to model alternative peer-review systems in a one-journal or systemic approach (Herron 2012; Allesina 2012). Most of these works have focused on specific questions about peer review, often reviewer behaviour, without considering the complete scientific publication system and without calibration with empirical data. However, to improve the peer-review system, we need to adopt a unified approach to both scientific publication and peer review that is more holistic and to use empirical data for calibrating models. Therefore, we have developed an ABM that we calibrated with empirical data pertaining to the biomedical domain (Kovanis et al. 2016a).

Here, our objective was to use an agent-based model to evaluate the efficiency of alternative peer-review systems currently implemented by some biomedical and general journals. We modified the ABM we previously developed to match the behaviour of these alternatives and compared their performance in terms of the base model. To our best knowledge, previous models focused mostly on microscopic behaviours; here we selected widely discussed systems requiring more macroscopic modifications to the ABM, which are largely understudied. Section ("Methods") contains a brief description of the base model for the conventional system, the alternative peer-review systems, their real-life examples and the changes we implemented in the sub-models of the conventional system. In "Results" section we present our results and our exploration of the parameter space. Finally, in "Discussion" section we discuss the implications of our results.

Methods

Overview

We used a previously developed ABM that was calibrated with empirical data and adopts a unified approach of scientific publication and peer review (Kovanis et al. 2016a). This ABM was structured in independently parameterized sub-models pertaining to the submission and peer-review process. Structural changes to some of these sub-models allowed us to model the alternative peer-review systems.

We compared five alternative systems of peer review discussed in the literature and to some extent already implemented by some journals and publishers: re-review opt-out, cascade peer review, portable peer review, crowdsourcing peer review, and immediate publication (Fig. 1). Their main characteristics and parameters are summarized in the Table 1.

Model for the conventional publication and peer-review system

Here we provide a brief description of our ABM of the conventional scientific publication and peer-review system. For a more detailed description, see Kovanis et al. (2016a).

We characterized N researchers by resources $R(t)$ and scientific level $S(t)$. The scientific level was defined as $S(t) = R(t) + S_b(t)$, where t the time step and $S_b(t)$ the sum of all the rewards that a researcher can receive to determine scientific level. The resources represent all the means that researchers have at their disposal for conducting research. The scientific level expresses a researcher's experience and capacity to conduct better research.

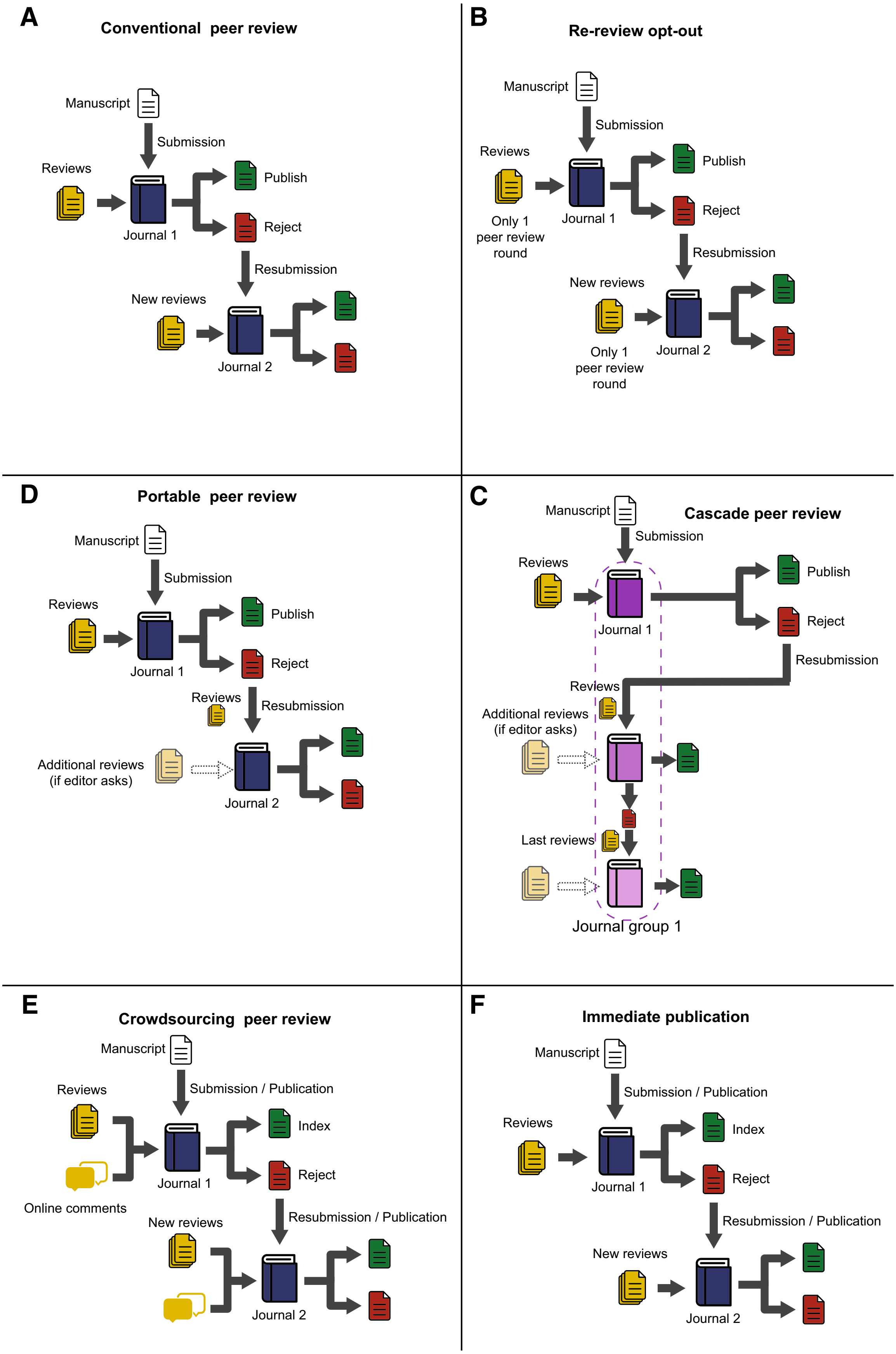


Fig. 1 Diagrams of the alternative peer-review systems

Table 1 Summary of the characteristics and parameters of the alternative peer-review systems

Peer-review systems	Main characteristics	Differences from the conventional system
Re-review opt-out	Only one round of peer review and revisions Acceptance or rejection depends on editor's evaluation of the revisions	*Evaluation of papers* Only 1 one round of peer review and revisions *Acceptance or rejection of papers* If paper is not rejected by the reviewers, then the editor evaluates (Q_e) its revised version (Q_{revised}) $Q_e \overset{\text{Uniformly drawn}}{\longleftarrow} [0.9Q_{\text{revised}}, 1.1Q_{\text{revised}}]$ Accepted only if the editor's evaluation is higher or equal to the acceptance threshold ($T_{\max}$) of the journal (j)
Cascade	Sharing of past reviews between journals belonging to the same group Resubmissions are allowed only in journals of the same publisher and of lower reputation	*Journals* Each journal belongs to one of the 4 groups that shares reviews internally *Decision on whether to ask for new reviews or not* The journal receives a paper of scientific value Q, its past reviews (Q_r) and the editor issues an evaluation (Q_e) If $Q \geq T_{\max}$ the paper is immediately accepted If $\frac{\lvert Q_e - Q_r \rvert}{Q_r} \leq 0.1$ the authors revise the paper and then the editor re-evaluates it and decides on acceptance or not If $\frac{\lvert Q_e - Q_r \rvert}{Q_r} > 0.1$ the editor asks for new reviews *Resubmission probability* The probability of resubmission (P_{res}) depends on whether the number of submissions (N_{sub}) is higher than in the conventional system $P_{\text{res}} = 0.88^{(N_{\text{sub}}-1)/2}$ instead of $P_{\text{res}} = 0.88^{(N_{\text{sub}}-1)}$ *Journal to resubmit* Randomly selected among the next 5 journals of lower reputation (belonging to the same group)
Portable	Sharing of past reviews between journals	*Decision on whether to ask for new reviews or not* Same as in the cascade system *Resubmission probability* Same as in the cascade system

Table 1 continued

Peer-review systems	Main characteristics	Differences from the conventional system
Crowdsourcing	Publication as "discussion papers" upon submission Editor takes into account possible online comments	*Initial scientific information (* $\mathrm{SI}_{\mathrm{init}} = \mathrm{AR}_j \times Q$*)* New submissions release initial scientific information depending on their scientific value (Q) and the journal (AR_j) *Evaluation of papers* Papers are evaluated by invited reviewers (N_R) and by a certain number of online commenters equal to $\frac{\mathrm{SI}_{\mathrm{init}}}{\mathrm{mean}(\mathrm{SI}_{\mathrm{total}})^2}$ mean ($\mathrm{SI}_{\mathrm{total}}$) is the average initial scientific information for all submissions in a time step *Acceptance or rejection of papers:* The editor evaluates the papers using the mean evaluation value of all the online comments (Q_{online}) and the evaluation of the invited reviewers (Q_{invited}) $Q_r = \frac{Q_{\mathrm{online}} + N_R Q_{\mathrm{invited}}}{N_R + 1}$ If the paper receives no online comments, then $Q_r = Q_{\mathrm{invited}}$ *Final scientific information (* $\mathrm{SI} = \mathrm{IF}_j \times Q_F$*)* All published papers release the rest of their scientific information ($\mathrm{SI} - \mathrm{SI}_{\mathrm{init}}$) at the time of acceptance For papers rejected and not resubmitted, 80% of their $\mathrm{SI}_{\mathrm{init}}$ is removed from the system
Immediate publication	Publication as "discussion papers" upon submission	*Initial scientific information* Same as in the crowdsourcing system *Final scientific information* Same as in the crowdsourcing system

Manuscripts were characterized by an intrinsic quality score (Q score), which serves as a proxy for their intrinsic scientific value but also their disruptive, innovative, or controversial nature as well as quality of reporting. At each time step, N_s randomly selected researchers submitted their paper. At the time of submission (t_s) of their paper, authors would lose an amount of resources R_{inv} associated with the conduct of the research reported in that paper with $0.2R(t_s) \leq R_{\mathrm{inv}} \leq 0.7R(t_s)$. Each paper had an initial expected quality E_Q defined as:

$$E_Q = 0.8\frac{0.1R_{\mathrm{inv}}}{0.1R_{\mathrm{inv}} + 1} + 0.2\frac{0.01S(t_s)}{0.01S(t_s) + 1}$$

The weights were chosen to represent the greater contribution of invested resources to the scientific level and to not allow the magnitude of $S(t_s)$ to surpass the final E_Q value. The Q score was drawn from a normal distribution $Q \sim N(E_Q,\ 0.1\ E_Q)$. This score determines how a researcher chooses a target journal and drives in-house and external peer-review assessments.

We characterized J journals by 3 state variables: a reputation value (we used rescaled impact factors) and by related rejection or acceptance thresholds, $T^j_{\min} < T^j_{\max}, j = 1, \ldots, J$.

We assumed that authors had a general knowledge of journal standards and, given the Q score, would try to obtain the most recognition from their work. Hence, the journal for the first submission was chosen at random among those with T^j_{min} within the asymmetrical range $Q - 0.45\varepsilon \leq T^j_{\min} \leq Q + 0.55\varepsilon$, where $\varepsilon \sim 2 \times N\left(\frac{Q}{5}, \frac{Q}{20}\right)$. This process resulted in a slight trend of high targeting in every first submission.

We drew the editor's assessment of the manuscript Q_e from a uniform distribution over $[0.9Q;\ 1.1Q]$. If $Q_e < T^j_{\min}$, the manuscript could be rejected without external peer review. If $Q_e \geq T^j_{\min}$, the manuscript was sent for external peer review to 2 or 3 reviewers. The reviewers' assessments were defined as $Q_r \sim N(Q - c,\ r \times Q)$, where r was a random error and c measured the competitiveness of the reviewer. We defined $r = r_r + r_j - r_Q$, where r_r is the reviewing error, r_j the journal error and r_Q the score error. With 65% probability, we set $r_t = 0.1$; with 12%, $r_t = 0.05$; and with 13%, $r_t = 0.01$. We drew r_j randomly from a uniform distribution over [0; 0.15], where $r_j = 0$ corresponded to the highest reputation journal and $r_j = 0.15$ to the lowest. Finally, $r_Q = 0.05 \times Q$. We assumed that a competitive behavior would occur more often for journals with higher reputation. The probability of appearance ranged uniformly from 10 to 66%, where c was drawn randomly from a uniform distribution over [0.01; 0.05].

We randomly selected one of the reviewers' evaluations as a proxy of the editor's opinion. If $Q_r \geq T_{\max}$, the manuscript was accepted and if $Q_r \leq T_{\min}$, it was rejected. When $T_{\min} \leq Q_r < T_{\max}$, the author was asked to revise the manuscript before a second round of peer review. In the latter case, the author invested an extra amount of resources $R_{\text{imp}} \sim N\left(\frac{8}{60}, \frac{1}{60}\right) \times (R - R_{\text{inv}})$. The cumulative amount of invested resources was used to derive a new Q score as before. The manuscript was re-evaluated and accepted only if $Q_r \geq T_{\max}$. The probability of resubmission P_{res} after a rejection decreased with increasing number of resubmissions r, $P_{\text{res}} = 0.88^{r-1}$. After the first rejection, the authors would target journals of lower reputation. Thus, they randomly selected journals in the (symmetrical this time) range $0.22Q - 0.5\varepsilon \leq T^j_{\min} \leq 0.22Q + 0.5\varepsilon$, where Q is the initial score of the manuscript.

Resources and scientific level were updated at each time step. If an article is published, the author received a random reward $p \times \left(R_{\text{inv}} + \sum_i R^i_{\text{imp}}\right), \quad 0 \leq p \leq 0.5$, otherwise, the author would permanently lose all the resources invested. In case of publication, the author also received a reward for resources in scientific level. The scientific level of a reviewer was credited with a random reward between 0 and 0.001 every time the reviewer completed a review because of knowledge acquired from the paper. Moreover, at the end of each week, the researchers received an update to their resources and scientific levels randomly drawn between 0.1 and 1, which reflected an increase of the means to conduct research with time.

We assumed that when a paper was published, it released scientific information to the community $\text{SI} = \text{IF}_j \times Q_F$, where IF_j is the reputation value of the journal (j) that published it and Q_F is the Q score of the paper, after all revisions. Scientific information is a comparative variable and its purpose is to assess the effectiveness of a system in producing more papers of higher Q score and in disseminating them to the rest of the scientific community. The reputation value (IF) of a scientific journal is a proxy of the size of the community that will read the paper and the Q score a proxy of how much people who read the paper will benefit from it.

Re-review opt-out

The intent of this system, currently implemented by BMC Biology, is to shorten the time of peer review by allowing authors to opt out from a second round of reviews. Thus, authors with a paper judged publishable with major revisions by the reviewers can choose whether they want their manuscript to be evaluated by the editor only or again by the reviewers after revising it (Robertson 2013).

We chose to model a maximum implementation of this intervention so that authors would always choose to opt out from a second round and therefore all decisions for every submission would be made after at most one peer-review round. For papers undergoing peer review, the authors always revised, and then the editor made an assessment (Q_e) of the revised manuscript from a uniform distribution between 0.9 Q_{revised} and 1.1 Q_{revised}. With $Q_e \geq T^j_{\max}$, a paper was accepted; otherwise it was rejected. All other processes were handled as in the conventional system.

Cascade peer review

When papers are rejected, their authors usually revise them and resubmit to other journals for publishing. In the conventional peer-review system, this implies that the same manuscripts will be reviewed multiple times and their publication can be seriously delayed. To avoid this situation, some publishers have decided to share reviews for rejected manuscripts among the journals they manage, thus avoiding redundant reviews and shortening the evaluation time. Such publishers include Nature Publishing Group, JAMA, BioMed Central and British Medical Journal (Walker and Rocha da Silva 2015a; Cals et al. 2013; Van Noorden 2013).

We randomly allocated 105 journals of various reputation value to one of four arbitrary publisher groups. We assumed that every journal belonged to one of these groups. Each journal was allocated to one of the publisher groups by using a categorical distribution with parameters (probability of belonging to each group) drawn from a normal distribution $\sim N(0.25, 0.025)$ for the three first groups, with the remaining ones allocated to the fourth.

When a paper was rejected, the editor proposed that the author send it to journals of the same network but of lower reputation. We assumed that if authors decided to resubmit, then they never rejected this proposal. Then, one of the next five journals of lower reputation value (of the same network) was randomly selected and the manuscript was resubmitted to it, along with the last evaluation value (Q_r).

The new editor immediately accepted the resubmitted paper without asking for further reviews if $Q_r \geq T_{\max}$; otherwise, the editor asked for revisions if $\frac{|Q_e - Q_r|}{Q_r} \leq 0.1$, where Q_e is the editor's assessment of the manuscript (drawn uniformly from between $0.9Q$ and $1.1Q$). Then the editor re-assessed the paper and decided whether to accept or reject it. Papers rejected were more likely to be resubmitted in this system than in the conventional system; thus the probability of resubmission was modified as $P_{\text{res}} = 0.88^{(N_{\text{sub}}-1)/2}$. Authors cascaded their submissions always using the last reviews they obtained. With $\frac{|Q_e - Q_r|}{Q_r} > 0.1$, the editor asked for new reviews and the submission was handled as in the base model.

Portable peer review

In this system, the authors resubmit their rejected manuscripts along with the reviews they received from their last peer-reviewed submission (if any). In contrast to the cascade system, the journals were not organized in groups and thus the authors sent their previous reviews to any of the journals they would be resubmitting to as in the conventional system. Based on the same rule as in the cascade system, editors could choose to ask for new reviews or revisions before deciding on acceptance or rejection.

Crowdsourcing peer review (Immediate publication with online and invited reviews)

Crowdsourcing online reviews is implemented in part by various journals such as F1000Research, Philica and the Semantic Web Journal. The purpose of this system is the immediate release of scientific information and the more accurate evaluation of papers because of any additional online comments or reviews. The journal of Atmospheric Chemistry and Physics (ACP) is also a well-known example of the use of such a system. Papers submitted to ACP pass a quick editorial pre-screening and are almost immediately published, following their submission, in the journal's website as "discussion papers". A published paper is then assigned external peer reviewers. The peer reviewers start an online discussion with the authors and other interested members of the scientific community. After a fixed number of weeks, the discussion stops and the authors revise the paper and resubmit it for publication (Walker and Rocha da Silva 2015a; Pöschl 2012; Journal 2015; Hunter 2012).

In our approach, papers were subject to traditional editorial assessment instead of a quick editorial pre-screening. This discussion did not have any pre-specified time limit and the rejected manuscripts could be left on the journal's webpage or resubmitted to another journal.

Each manuscript that passed the conventional in-house review stage was immediately published, along with a call for online reviews/comments and the traditional invitation to two or three external reviewers selected by the journal's editor. Every fresh submission released an initial amount of scientific information, $\mathrm{SI}_{\mathrm{init}} = \mathrm{AR}_j \times \mathrm{Q}$, where AR_j represents the reputation of the "discussion papers" section of the journal (j). We obtained AR_j from the original simulations of the conventional system, and it is equal to the acceptance rate of papers, after the editorial screening process. We assumed that a publication attracted a number of online reviewers equal to $\frac{\mathrm{SI}_{\mathrm{init}}}{\mathrm{mean}(\mathrm{SI}_{\mathrm{total}})^2}$, rounded to the nearest integer, where $\mathrm{mean}(\mathrm{SI}_{\mathrm{total}})$ is the average $\mathrm{SI}_{\mathrm{init}}$ of all papers submitted at each time step ($\mathrm{SI}_{\mathrm{total}}$ represents the distribution of $\mathrm{SI}_{\mathrm{init}}$ values at a time step).

The online commenters evaluated the paper in the same way as the normal reviewers. The editor averaged the scores of the online commenters (Q_{online}) and randomly selected one of the invited reviewers' scores (Q_{invited}), as in the conventional system to make a decision (Q_{r}). We assumed that editors took more into account comments from reviewers they invited than uninvited reviewers, thus $Q_r = \frac{Q_{\mathrm{online}} + nQ_{\mathrm{invited}}}{n+1}$, where n is the number of invited reviewers. Thus, the more online reviewers, the greater the chance a paper was more accurately evaluated. If the paper did not attract any online comments, then $Q_r = Q_{\mathrm{invited}}$.

With $Q_r \geq T_{\max}$, the paper was revised once, considered indexed in the bibliographical databases (Web of science, MEDLINE etc.) and included as a part of the next issue of the journal, thus releasing the rest of its scientific information at the time of indexation. With $Q_r < T_{\max}$, the authors decided to resubmit based on $P_{\mathrm{res}} = 0.88^{N_{\mathrm{sub}}-1}$ or leave their paper unindexed on the webpage of the journal. In the latter case, the paper would still release some scientific information because it can be found online but less so because it will be hidden in the journal's website. Thus, subtracting an amount from the total scientific information the scientific community had already accumulated (because of the paper's higher visibility as a "discussion" paper), the manuscript's final scientific information becomes $SI = 0.2\ SI_{init}$.

Immediate publication

In the system of immediate publication, papers are immediately available to the readers as "discussion papers" before they are peer reviewed via the webpage of the journal. This system is similar to the crowdsourcing system ("Crowdsourcing peer review (Immediate publication with online and invited reviews)" section) but without assuming that editors would take into consideration any online reviews or comments.

Implementation and system comparison

We programmed the models by using MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, MA, USA) with a total number of researchers $N = 25{,}000$, total number of journals $J = 105$ and weekly submissions drawn from a normal distribution $\sim N\,(850, 85)$ (each simulation week is 1 time-step). We ran the simulations for 10 years, with a burn-in period of 1 year for the initialization of the model. All main results were averaged over 100 simulation runs. Code is available at http://www.clinicalepidemio.fr/peerreview_alternative_systems/.

We defined three different types of outcomes to compare all alternative systems with the conventional system; peer-review efficiency, reviewer effort and scientific dissemination. Peer-review efficiency corresponded to the double purpose of peer review. We measured it by using the separation of the Q score distributions of the published and unpublished papers and the relative improvement in average Q score for all papers after revision as compared to that for the first submission. We used the Hellinger distance as a quantifying measure of the overlap between two distributions: the higher the Hellinger distance, the less the overlap (Nikulin 2001). We measured reviewer effort by using the total time reviewers devoted to peer review in a year. We obtained this outcome in hours from our simulations and transformed it into working years per year with the following equation:

$$\text{time spent in peer review} = \frac{\text{hours devoted to peer review/work hours}}{\text{year} - \text{weekends} - \text{holidays}}$$

where work hours = 8 h per day, year = 365 days, weekends = 104 days and holidays = 25.3 days (average paid holidays in 21 OECD countries) (Ray and Schmitt 2007). Finally, we measured scientific dissemination by using the number of annual publications, the median weeks between first submission of a paper and the final decision, the average Q score for all papers and the average weekly release of scientific information. For estimating the two time-related measures, we used the respective distributions from an international survey of 4000 participants (Mulligan et al. 2013).

Finally, we considered that a peer-review system was beneficial if it improved any of the outcomes without deteriorating the peer-review efficiency and more efficient than the conventional if it improved all types of outcomes.

Sensitivity analyses

We performed sensitivity analyses of two of the alternative peer-review systems: cascade and crowdsourcing. We excluded the re-review opt-out, portable and immediate publication systems because the first is already at its maximum configuration and cannot realistically be improved in our ABM and the second and third can be considered special cases of the cascade and crowdsourcing systems, respectively. These analyses focused on identifying the effect of different configurations of the cascade and crowdsourcing systems on their outputs. All sensitivity analyses were averaged over 10 simulation runs.

Exploring the parameter space for the cascade peer-review system

In the main version of the cascade system, with initialized $N_g = 4$ journal groups, the editor asks for new reviews or not based on $\frac{|Q_e - Q_r|}{Q_r} \leq \alpha$, where $\alpha = 0.1$, and the probability that an author accepts the editor's proposal is $P_{\text{cas}} = 1.0$. We explored the parameter space by varying these three parameters one at a time while keeping the other two the same as in the main version of the cascade system. We ran the cascade system for $N_g = 2, 3$ and 5, for $\alpha = 0.0$ and 1.0 and for $P_{\text{cas}} = 0.7, 0.8$ and 0.9. The cases with $\alpha = 0.0$ and 1.0 represent those for which all and none of the resubmitted papers receive new peer review, respectively.

Effect of the editor's decision and online comments with the crowdsourcing system

We explored different assumptions on how editors decide on acceptance or rejection of a paper and how the online comments affect the system overall. Here we explored the cases in which all papers received 1, 5 and 20 comments. Moreover, we simulated when editors averaged all reviews, online and invited, and when they chose at random one of the online or invited peer reviews to represent their decision. The last two cases assumed a mechanism of attracting online comments identical to the main version of the system.

Results

Peer-review efficiency (Table 2)

Only the cascade and the crowdsourcing peer-review systems outperformed the conventional system for both outcomes. Their performance was similar in terms of separation of Q score distributions; however, the cascade system outperformed both the conventional and crowdsourcing systems in terms of improving the Q scores of submitted papers and the average weekly release of scientific information. The immediate publication system performed almost identically to the conventional system, and the portable and re-review opt-out systems failed to match that of the conventional system in one and two of the measures, respectively.

Reviewer effort (Table 2)

The best-performing systems were the cascade and portable peer-review systems. They had the highest deviation from the conventional system performance. The systems took about

Table 2 Values of all outcome measures of all peer-review systems implemented

Outcome measures	Conventional	Re-review opt-out	Cascade	Portable	Crowdsourcing	Immediate publication
Peer-review efficiency						
Separation of the Q score distributions of the published and unpublished papers (HD)	0.433 ± 0.003	0.312 ± 0.105 (−28.0%)	0.452 ± 0.008 (+4.4%)	0.414 ± 0.004 (−4.2%)	0.448 ± 0.003 (+3.5%)	0.447 ± 0.003 (+3.2%)
Relative improvement of Q score of papers after revisions	5.51 ± 0.03%	3.32 ± 0.03% (−39.6%)	6.27 ± 0.04% (+13.9%)	5.96 ± 0.02% (+8.1%)	5.66 ± 0.03% (+2.7%)	5.65 ± 0.03% (+2.6%)
Reviewer effort						
Time spent in peer review (work years/year)	971 ± 12	684 ± 9 (−20.5%)	360 ± 6 (−62.9%)	347 ± 6 (−64.3%)	992 ± 12 (+2.1%)	995 ± 14 (+15.7%)
Scientific dissemination						
Annual no. of publications	31,425 ± 164	33,743 ± 165 (+7.4%)	29,757 ± 389 (−5.3%)	33,614 ± 184 (+7.0%)	31,143 ± 158 (−0.9%)	31,199 ± 178 (−0.7%)
Time between first submission and final decision (weeks), median	15	14 (−6.7%)	7.9 ± 0.2 (−47.3%)	8 (−46.7%)	15 (0.0%)	15 (0.0%)
Q score of papers	0.8182 ± 0.0007	0.8034 ± 0.0008 (−1.8%)	0.8254 ± 0.0023 (+0.9%)	0.8502 ± 0.0006 (+3.9%)	0.8229 ± 0.0006 (+0.6%)	0.8229 ± 0.0006 (+0.6%)
Release of scientific information (per week)	27.4 ± 0.2	28.0 ± 0.3 (+4.3%)	37.4 ± 0.6 (+36.6%)	30.2 ± 0.2 (+10.2%)	34.4 ± 0.3 (+25.7%)	34.5 ± 0.3 (+26.0%)

Data are mean ± SD unless indicated from 100 simulation runs. HD, Hellinger distance

Parentheses indicate the relative difference for each outcome to the conventional system

For median weeks between first and last submission and final decision, SD = 0

60% less time for review of all submissions. The re-review opt-out system was also beneficial in terms of total time devoted to peer review, which was 20.5% less than in the conventional system. The immediate publication and crowdsourcing peer-review systems performed slightly worse than the conventional system.

Scientific dissemination (Table 2)

The most beneficial systems were the cascade and portable peer-review systems. They both shortened the time to publication by about 47% and increased the average weekly release of scientific information by 36.6 and 10.2%, respectively. The average Q score for all articles was also higher, by 0.9 and 3.9%. Moreover, the portable system published 7.0% papers more than the conventional system, but the cascade system 5.3% less. The re-review opt-out system was also beneficial in terms of papers published per year (7.4% higher), median time to publication (6.7% less) and average weekly release of scientific information (2.6% higher). Finally, the crowdsourcing and immediate publication systems differed from the conventional only in terms of release of scientific information, which was 26.0% higher for both systems.

Overall evaluation of the systems

We considered that a system could be more efficient than the conventional system only if it improved all types of outcome measures and beneficial if it improved at least one outcome without deteriorating peer-review efficiency. Among all alternatives, only the cascade system was more efficient than the conventional system. Moreover, the crowdsourcing and immediate-publication systems were beneficial in terms of scientific dissemination. The re-review opt-out, while advantageous in some of the measures, severely deteriorated peer-review efficiency. Finally, the portable peer review was advantageous in terms of almost all outcome measures but failed to match at least the performance of the conventional system in terms of separation of Q score distributions.

Sensitivity analyses

Exploration of the parameter space for the cascade peer-review system (Table 3)

Most of the different configurations of the cascade system surpassed or matched the performance of the conventional system in peer-review efficiency (apart from $P_{cas} \leq 0.80$) and reviewer effort measures and all outperformed the system in median time to the final decision and release of scientific information. However, the number of published papers was lower for all alternative systems than the conventional system. The best-performing configuration was the one with $\alpha = 1.0$, whereby the editors never asked for new reviews on resubmitted papers.

Effect of the editor's decision and online comments on the crowdsourcing system (Table 4)

All the different configurations of the crowdsourcing system matched or over-performed the conventional system in terms of peer-review efficiency and weekly release of scientific

Table 3 Values for all outcome measures for all configurations of the cascade system

Outcome measures	N_g =			α =		P_{cas} =		
	2	3	5	0.0	1.0	0.9	0.8	0.7
Peer-review efficiency								
Separation of the Q score distribution (HD)	0.444 ± 0.005 (−1.7%)	0.449 ± 0.008 (−0.6%)	0.458 ± 0.004 (+1.4%)	0.451 ± 0.007 (+0.2%)	0.456 ± 0.008 (+0.9%)	0.429 ± 0.004 (−5.1%)	0.411 ± 0.004 (−9.0%)	0.38 ± 0.05 (−15.9%)
Relative improvement of Q score	6.41 ± 0.02% (+2.2%)	6.35 ± 0.03% (+1.3%)	6.21 ± 0.04% (−1.0%)	5.95 ± 0.04% (−5.2%)	6.24 ± 0.04% (−0.4%)	6.09 ± 0.04% (−2.8%)	5.90 ± 0.03% (−5.8%)	5.73 ± 0.03% (−8.6%)
Reviewer effort								
Time spent in peer review (work years/year)	359 ± 8 (+0.2%)	358 ± 5 (+0.4%)	358 ± 6 (+0.4%)	540 ± 10 (+50.1%)	211 ± 5 (−41.0%)	424 ± 5 (+17.9%)	491 ± 7 (+36.6%)	557 ± 7 (+54.9%)
Scientific dissemination								
Annual no. of publications	27,422 ± 195 (−7.9%)	28,724 ± 368 (−3.5%)	30,182 ± 320 (+1.4%)	29,286 ± 258 (−1.6%)	29,794 ± 455 (+0.1%)	29,216 ± 354 (−1.8%)	28,409 ± 187 (−4.5%)	27,602 ± 150 (−7.2%)
Time between first submission and final decision (weeks), median	7 (−11.5%)	7 (−11.5%)	8 (+1.2%)	8 (+1.2%)	7 (−11.5%)	8 (+1.2%)	8 (+1.2%)	8 (+1.2%)
Q score of papers	0.810 ± 0.001 (−1.9%)	0.820 ± 0.003 (−0.7%)	0.8272 ± 0.0018 (+0.2%)	0.8197 ± 0.0011 (−0.7%)	0.8268 ± 0.0023 (+0.2%)	0.8010 ± 0.0011 (−2.9%)	0.8102 ± 0.0011 (−1.8%)	0.8293 ± 0.0023 (+0.5%)

Table 3 continued

Outcome measures	$N_g =$			$\alpha =$		$P_{\text{cas}} =$		
	2	3	5	0.0	1.0	0.9	0.8	0.7
Release of scientific information (per week)	41.8 ± 0.3 (+11.6%)	39.4 ± 0.7 (+5.2%)	36.2 ± 0.4 (−3.3%)	36.6 ± 0.3 (−2.2%)	37.5 ± 0.7 (+0.2%)	34.7 ± 0.5 (−7.3%)	32.4 ± 0.4 (−13.5%)	30.1 ± 0.3 (−19.6%)

Data are mean ± SD unless indicated from 100 simulation runs. HD, Hellinger distance

Parentheses indicate the relative difference for each outcome to the conventional system

For median weeks between first and last submission and final decision, SD = 0

Table 4 Values for all the outcome measures for all different configurations of the crowdsourcing system

Outcome measures	Crowd-sourcing	Immediate publication	1 online comment	5 online comments	20 online comments	Average of all comments and reviews	Randomly select one comment or review
Peer-review efficiency							
Separation of the Q score distribution (HD)	0.448 ± 0.003	0.447 ± 0.003 (−0.2%)	0.446 ± 0.002 (−0.4%)	0.448 ± 0.002 (0%)	0.449 ± 0.003 (+0.2%)	0.468 ± 0.002 (+4.5%)	0.434 ± 0.001 (−3.1%)
Relative improvement of Q score	5.66 ± 0.05%	5.65 ± 0.03% (−0.02%)	5.67 ± 0.05% (+0.2%)	5.65 ± 0.05% (+0.3%)	5.67 ± 0.05% (+0.2%)	5.82 ± 0.01% (+2.9%)	5.51 ± 0.03% (−2.6%)
Reviewer effort							
Time spent in peer review (work years/year)	991 ± 12	995 ± 14 (+0.4%)	990 ± 12 (−0.2%)	989 ± 17 (−0.3%)	1000 ± 9 (+0.8%)	1009 ± 6 (+1.7%)	986 ± 7 (−0.6%)
Scientific dissemination							
Annual no. of publications	31,143 ± 158	31,199 ± 178 (+0.2%)	31,161 ± 152 (+0.1%)	31,052 ± 114 (−0.3%)	31,129 ± 156 (0.0%)	30,682 ± 188 (−1.5%)	31,299 ± 157 (+0.5%)
Time between first submission and final decision (weeks), median	15	15 (0.0%)	15 (0.0%)	15 (0.0%)	15 (0.0%)	15 (0.0%)	15 (0.0%)
Q score of papers	0.8229 ± 0.0006	0.8229 ± 0.0006 (0.0%)	0.8229 ± 0.0005 (0.0%)	0.8226 ± 0.0004 (0.0%)	0.8225 ± 0.0005 (0.0%)	0.8231 ± 0.0004 (0.0%)	0.8225 ± 0.0007 (0.0%)
Release of scientific information (per week)	34.4 ± 0.3	34.5 ± 0.3 (+0.3%)	34.5 ± 0.4 (+0.3%)	34.4 ± 0.2 (0.0%)	34.5 ± 0.2 (+0.3%)	34.2 ± 0.2 (−0.6%)	34.5 ± 0.3 (+0.3%)

Data are mean ± SD unless indicated from 100 simulation runs. HD, Hellinger distance

Parentheses indicate the relative difference for each outcome to the conventional system

For median weeks between first and last submission and final decision, SD = 0

information but without providing any advantage in reviewer effort and the other scientific dissemination measures.

Discussion

We implemented several structural modifications to an original ABM of the conventional scientific publication and peer-review system and modelled five alternative peer review systems to compare their performance and relative efficiency in terms of certain outcomes. In our simulations, cascade peer review was the only alternative more efficient than the conventional one. Cascade peer review is based on the trade-off between agreeing to submit and publish in journals of lower reputation and publishing faster than usual. Under our assumptions, the number of total annual publications slightly decreased by about 5.3%, but the total time reviewers devoted to peer review decreased by 62.9% and the total time from first submission to final decision decreased by 47.3%. These results came without deterioration in the peer-review efficiency measures and even with some improvement. Most notably, this system increased the average weekly release of scientific information by 36.6%, outperforming even the two systems with immediate publication.

We did not reallocate the time researchers saved from peer review to more resources available for research, and thus we might have underestimated the advantages of both cascade and portable peer-review. For example, this reallocation of resources could lead to higher-quality review reports because reviewers are not overburdened with the task. This reallocation could also help reviewers in the re-review opt-out system raise their overall screening ability. Moreover, this time not spent on peer review could also be reallocated to more resources for research and thus raise the average Q score of manuscripts. Still, the systems with immediate publication release fast new information, which is reallocated to the authors as a small bonus in scientific level. However, the fact that research can be communicated very fast is something that in reality can benefit the world way more than our simulations can portray.

From the similarities and differences between the results of the two review-sharing systems, we can see how their microscopic assumptions affect the macroscopic picture. First, only their review-sharing aspect led to results of the time metrics decreasing in comparison to the conventional system. Cascading submissions to journals of lower impact factor did not affect the speed of publication and did not provide any personal advantages to authors. This occurred in cascade peer review because any paper of low Q score submitted in a journal network that did not include journals of very low standards would most of the time be rejected. However, cascading submissions provided some overall advantages by better separating the Q-score distributions because of the rejection of papers that would have been published in the portable system.

On further investigation of the configurations of the cascade system, its main configuration was not the only one providing these advantages. The best-performing configuration was the one in which the editors never asked for new reviews for any resubmitted paper. This configuration required 41% less time reviewers devoted to peer review than with the main configuration and one week less time to a final decision. This result is important because if papers were evaluated only once, they would require about 78% less time from reviewers than what they do now. However, in real life this rule could be potentially abused by reviewers with, for example, competitive motives resulting in manuscripts with unfair reviews carried forward along resubmissions. The passing of reviews should therefore not be implemented strictly and editors should always be able to ask for additional reviews if reviews appear overly negative. Moreover, we explored how the number of the journal groups affected the results. This kind of exploration essentially affected the gap in impact factor for journals between resubmissions of rejected manuscripts. The differences in number of groups of journals did not affect the results greatly, with the exception that for two or three groups, it took one week less to a final decision. To be more efficient than the conventional, the cascade system requires that the authors accept more than 90% of the time the editor's proposal to send their paper to a journal of lower impact factor along with the reviews.

The system of portable peer review was modelled exactly as the cascade system, with the only exception that authors were not restricted by journal groups when resubmitting. Our results suggest that this system is also beneficial, almost as much as the cascade system. However, the 4.2% decrease in separation between the Q score distributions of the published and unpublished papers is undesirable. The portable system, despite its small disadvantage in separation of Q score distributions, might be easier to implement in real-life because it provides authors with more freedom to resubmit.

The system of crowdsourcing online reviews was beneficial but not more efficient than the conventional system. Simply by implementing its immediate publication version, without online reviews, increased the release of scientific information by 25.7%. Then, introducing online reviews to the system increased both peer-review efficiency outcome measures because of more correct evaluation of papers due to the fact that editors obtain more reviews. Online reviews are rarely as detailed as those from invited reviewers and thus we assumed that the editor assigned them lower weight than the invited ones. Moreover, since the results for only one online review per paper are the same as those for 5 or 20, averaging all the online reviews did not affect our outputs. Finally, in the extreme case, in which all online reviews were as detailed as the invited ones and all would be equally averaged, the system clearly managed to separate the Q score distributions better than any other. However, when we randomly selected one review, the system matched the behaviour of the conventional system.

The system of re-review opt-out is conceptually easy to implement however failed to at least match the performance of the conventional system on the two peer-review efficiency measures but improved on almost all the remaining outcome measures. In our implementation, we substituted the second round of revisions by the reviewers with an editorial evaluation. Thus, a real-life experiment and extra modelling efforts are needed to validate whether we obtained these results for the two outcome measures due to our modelling assumptions, which give high importance to the second peer-review round, or because this system is really less efficient than the conventional one.

A limitation of our simulations is that to our best knowledge no data currently exists for any of the five implemented alternative peer-review systems. For this reason, we had to obtain results by comparing the alternatives with the conventional system. However, these alternatives are not yet fully implemented and much of the relevant data are not even generated to date. A second limitation is that our results are likely affected by our assumptions and choices, more than the general idea behind these alternative systems. In general, we tried to adopt the most reasonable implementations of these alternatives in our main simulations and to test their limits and our assumptions by further exploring the parameter space for the two most important systems. Finally, our outcome measures were based on variables that are abstract in how we measured them. In theory, papers have a Q score that can act as a proxy of their novelty and correctness, for instance, and information is disseminated when journals publish new papers. However, because we lack a universally agreed-upon method and variables that measures these values, we needed to create them to help inform our decisions. These assumptions can only be proven or disproven after real-life experiments.

Conclusions

We compared the efficiency of five alternative peer-review systems to the conventional system by using an ABM approach. Only the cascade system was more efficient than the conventional system in all three types of outcomes. The portable system closely matched the cascade system's performance and was more efficient than the conventional system in all but one measure. Moreover, all the configurations of the crowdsourcing system were beneficial and managed to match or improve the peer-review efficiency and scientific information measures but without any important change in the other measures. Finally, we recommend prioritizing a system of review sharing to create a sustainable scientific publication and peer-review system.

Acknowledgements This work was partially supported by the COST Action TD1306 "New frontiers of peer review (www.peere.org)". Michail Kovanis is the recipient of a Ph.D. grant from Sorbonne Paris Cité. We thank Elise Diard (Cochrane France) for her help with the diagrams and Laura Smales (BioMedEditing, Toronto, ON) for copyediting the manuscript.

References

Allesina, S. (2012). Modeling peer review: An agent-based approach. *Ideas in Ecology and Evolution, 5*(2), 27–35. doi:10.4033/iee.2012.5b.8.f.

Arns, M. (2014). Open access is tiring out peer reviewers. *Nature, 515*(7528), 467. doi:10.1038/515467a.

Bianchi, F., & Squazzoni, F. (2015). Is three better than one? Simulating the effect of reviewer selection and behavior on the quality and efficiency of peer review. In L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, & M. D. Rossetti (Eds.), *Proceedings of the 2015 Winter Simulation Conference* (pp. 4081–4089). NJ: IEEE Press, Piscatawaw. doi:10.1109/WSC.2015.7408561.

Bohannon, J. (2013). Who's afraid of peer review? *Science, 342*(6154), 60–65. doi:10.1126/science.342.6154.60.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences, 99*(suppl 3), 7280–7287. doi:10.1073/pnas.082080899.

Bruce, R., Chauvin, A., Trinquart, L., Ravaud, P., & Boutron, I. (2016). Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. *BMC Medicine*. doi:10.1186/s12916-016-0631-5.

Cals, J. W., Mallen, C. D., Glynn, L. G., & Kotz, D. (2013). Should authors submit previous peer-review reports when submitting research papers? Views of general medical journal editors. *The Annals of Family Medicine, 11*(2), 179–181. doi:10.1370/afm.1448.

Galea, S., Riddle, M., & Kaplan, G. A. (2010). Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology, 39*(1), 97–106. doi:10.1093/ije/dyp296.

Gura, T. (2002). Scientific publishing: Peer review, unmasked. *Nature, 416*(6878), 258–260. doi:10.1038/416258a.

Herron, D. (2012). Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review. *Surgical Endoscopy, 26*(8), 2275–2280. doi:10.1007/s00464-012-2171-1.

Hopewell, S., Collins, G. S., Boutron, I., Yu, L.-M., Cook, J., Shanyinde, M., et al. (2014). Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. *BMJ*. doi:10.1136/bmj.g4145.

Houry, D., Green, S., & Callaham, M. (2012). Does mentoring new peer reviewers improve review quality? A randomized trial. *BMC Medical Education, 12*(1), 83.

Hunter, J. (2012). Post-publication peer review: Opening up scientific conversation. *Frontiers in computational neuroscience, 6,* 63.

Jennings, C. (2006). Quality and value: the true purpose of peer review. What you can't measure, you can't manage: the need for quantitative indicators in peer review. *Nature*. doi:10.1038/nature05032.

Journal, S. W. (2015). For reviewers. Available at: http://www.semantic-web-journal.net/reviewers. Accessed October 11, 2016.

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016a). Complex systems approach to scientific publication and peer-review system: development of an agent-based model calibrated with empirical journal data. *Scientometrics, 106*(2), 695–715. doi:10.1007/s11192-015-1800-6.

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016b). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS ONE, 11*(11), e0166387. doi:10.1371/journal.pone.0166387.

Marshall, B. D., & Galea, S. (2015). Formalizing the role of agent-based modeling in causal inference and epidemiology. *American Journal of Epidemiology, 181*(2), 92–99. doi:10.1093/aje/kwu274.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology, 64*(1), 132–161. doi:10.1002/asi.22798.

Nicholas, D., Watkinson, A., Jamali, H. R., Herman, E., Tenopir, C., Volentine, R., et al. (2015). Peer review: Still king in the digital age. *Learned Publishing, 28*(1), 15–21. doi:10.1087/20150104.

Nikulin, M. S. (2001). Hellinger distance. In: *Encyclopedia of mathematics*.

Paolucci, M., & Grimaldo, F. (2014). Mechanism change in a simulation of peer review: From junk support to elitism. *Scientometrics, 99*(3), 663–688. doi:10.1007/s11192-014-1239-1.

Park, I. U., Peacey, M. W., & Munafo, M. R. (2014). Modelling the effects of subjective and objective decision making in scientific peer review. *Nature, 506*(7486), 93–96. doi:10.1038/nature12786.

Patel, J. (2014). Why training and specialization is needed for peer review: A case study of peer review for randomized controlled trials. *BMC Medicine*. doi:10.1186/s12916-014-0128-z.

Pöschl, U. (2012). Multi-stage open peer review: Scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. *Frontiers in computational neuroscience, 6,* 33.

Ray, R., & Schmitt, J. (2007). No-vacation nation USA—a comparison of leave and holiday in OECD countries. *IEEE Policy Brief, 3*.

Rennie, D. (2016). Let's make peer review scientific. *Nature, 535*(7610), 31–33. doi:10.1038/535031a.

Righi, S., & Takács, K. (2017). The miracle of peer review and development in science: An agent-based model. *Scientometrics*. doi:10.1007/s11192-017-2244-y.

Robertson, M. (2013). Q&A: Re-review opt-out and painless publishing. *BMC Biology, 11,* 18. doi:10.1186/1741-7007-11-18.

Sense About Science. (2012). *Peer review: The nuts and bolts*. London: Sense About Science.

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences, 112*(2), 360–365. doi:10.1073/pnas.1418218112.

Squazzoni, F., & Gandelli, C. (2013). Opening the black box of peer review: An agent-based model of scientist behaviour. *Journal of Artificial Societies and Social Simulation, 16*(2), 3.

Stahel, P. F., & Moore, E. E. (2014). Peer review for biomedical publications: We can improve the system. *BMC Medicine*. doi:10.1186/s12916-014-0179-1.

Thurner, S., & Hanel, R. (2011). Peer-review in a world with rational scientists: Toward selection of the average. *The European Physical Journal B, 84*(4), 707–711. doi:10.1140/epjb/e2011-20545-7.

Van Noorden, R. (2013). Company offers portable peer review. *Nature, 494*(7436), 161.

van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ, 318*(7175), 23–27. doi:10.1136/bmj.318.7175.23.

Vespignani, A. (2012). Modelling dynamical processes in complex socio-technical systems. *Nat Phys, 8*(1), 32–39. doi:10.1038/nphys2160.

Walker, R., & Rocha da Silva, P. (2015a). Emerging trends in peer review: A survey. *Frontiers in Neuroscience, 9,* 169. doi:10.3389/fnins.2015.00169.

Walker, R., & Rocha da Silva, P. (2015b). Emerging trends in peer review: A survey. *Frontiers Neurosci, 9,* 169. doi:10.3389/fnins.2015.00169.

Ware, M. (2013). *Peer review: An introduction and guide*. Bristol: Publishing Research Consortium.

General discussion

Summary of the results

In this PhD thesis, I created a mathematical model to evaluate the potential supply and demand for peer review as well as the imbalance in the effort by researchers. Following that, I developed an agent-based model of the scientific publication and peer-review systems, calibrated it with empirical data from the biomedical domain and modified it to mimic the behavior of alternative peer-review systems. I compared these systems with the results of the conventional to draw conclusions on whether any of them may be more efficient in terms of peer-review efficiency, reviewer effort and scientific dissemination. In all the projects, my modeling approach was mainly focused on macroscopic aspects of scientific publication and was informed with empirical data to reflect reality as much as possible.

My results challenge the dominant, but anecdotal, claim that there is a shortage of reviewers due to the increase of published articles. This work shows that for the last 26 years there was not a shortage in the potential supply of reviewers as compared to the demand. In fact, for 2015 the potential supply exceeded the demand for reviewers and reviews by 15% to 249%. However, there was an important imbalance in the peer-review effort, with 20% of researchers performing 69% - 94% of all reviews. I estimated the time reviewers devoted to peer review in

2015 to be 63.4 million hours, in which the top 5% of the reviewers was responsible for 18.9 million hours.

My agent-based modeling approach focused on evaluating whether five alternative systems minimized the reviewer effort, optimized scientific dissemination and kept peer-review efficiency at least to the levels of the conventional system. The two peer-review strategies, which allow sharing of reviews when rejected manuscripts are resubmitted, appeared to perform the best in all metrics. Systems of immediate publication of scientific manuscripts, with or without online comments, improved some dissemination metrics but provided no other important advantage. Finally, I showed that using only one round of peer review and revisions made the system less efficient in terms of peer-review efficiency, than it currently is.

My modeling approach is the first to combine the macroscopic characteristics of scientific publication with the microscopic characteristics of peer review, while being calibrated with empirical data. Most of the previous models describe peer review in the context of a single journal, thus not capturing the systemic aspect of it. Moreover, almost none of them was informed with empirical data, thus the results of the simulations were too abstract and possibly very far from reality. In my agent-based model, I tried to capture the complete complexity of the system and keep its results as close to reality as possible.

Limitations

My first project has certain limitations. First, the quantification of the potential supply and demand for peer review relied on assumptions on the values of parameters in which no empirical data were accessible. However, this only concerned a limited number of parameters, on which I performed extensive sensitivity analyses. Moreover, some of the data did not come directly from the publishers, but from second sources such as Publons and surveys. Even though this limits my study to some extent, data on the peer-review system is very scarce and all the sources I used were those of the highest quality available at that time. Another limitation of this study was that I did not consider subdomains of biomedicine and individual interactions between authors, reviewers and editors. However, my very objective was to study the only overall quantitative supply and demand. Subdomains and interactions in the micro-scale would be useful to be studied, but with a different modeling approach, such as agent-based modeling.

My agent-based model is also limited by the lack of empirical data for some of the parameters. However, for some parts of the scientific publication system, data were impossible to be obtained, either because they do not exist (the real Q score distribution) or because they are too difficult to be collected (the volume and resubmission patterns of unpublished manuscripts). Moreover, any biases coming from assumptions in the model may be carried out in the same way in the alternative systems and thus, likely they may cancel out. In addition, I was not able to test all possible configurations of those alternative peer-review systems I modeled. This would require excessive amounts of time, because the ways that an alternative system may be implemented are numerous. Therefore, I simulated only a limited set of their most reasonable configurations.

Impact and implications

Many times editors face difficulties in finding reviewers and, as the data from Publons suggests, some researchers are performing too many reviews per year. My results show that this problem does not arise from the fact that there are no available reviewers, but rather that there are too many who review too few papers or not any at all. Even though one might think that researchers who are not involved in the peer-review process are not qualified for the task, relevant research shows that this is not the case. It has been pointed out that editors are biased towards inviting less women and Chinese researchers, while young academics who are willing to review do not know how to be involved in the process (Taylor & Francis group 2016; Warne 2016; Primack et al. 2016; Culley 2017; Helmer et al. 2017). Moreover, editors many times lean too heavily on reviewers they personally know and trust. One solution that has been proposed is to assign the reviewing task to professional services, while others believe that it is the responsibility of editors to expand their reviewer pool by reaching out to those who review the least in order to relieve those who review the most (Davis 2016; Oransky & Marcus 2016; Osterath 2016).

A technical issue that contributes to this problem is the heterogeneity of journal databases of reviewers. Each scientific journal has its own database, which usually is not shared even between journals belonging to the same publisher or group. These databases may contain names of authors who have submitted a manuscript to the journal and of scientists that editors asked in the past to review a manuscript. The latter may be personal contacts of the editor or scientists that the editor knows they are working on a relevant field. Scientists unknown to editors or to the databases of their journals may not be contacted for peer review. Therefore, authors who submit a lot are more likely to be known by many databases and overburdened with invitations to review. In addition, editors may also prioritize papers they judge to be important to be sent for review by reviewers they know to be reliable. Thus, the rest of the papers may be prioritized lower and sent to reviewers less reliable or more unknown to the editor. This practice, if performed by a large number of editors, may also help explain the big imbalance in the distribution of reviewer effort (Davis 2016).

Even though decreasing the time devoted to peer review may not be top priority for the community, since the potential supply exceeds the demand, its order of magnitude is still enormous. One way to diminish it may be to disincentivize scientists from publishing a large amount of papers during their career and to incentivize them to publish fewer papers of potentially higher quality. The "publish or perish" culture in science seems to be one important cause for the influx in the number of published scientific papers during the last decades. Moreover, it may be a reason that the business model followed by predatory journals managed to survive. Changing this practice implies that the scientific community, or the funding agencies, need to agree on more objective measures of the quality of scientists' work than the amount of their publications and the impact factor of the journals that published them.

Another way may be to apply interventions to the system of scientific publications itself. For instance, interventions such as the sharing of reviews between journals when resubmitting a manuscript seem to be promising. My agent-based simulations provide an approach that may be used for testing much more alternative interventions on peer review, than those presented in this project. Thus, one may devise many interventions, simulate them and compare their results, all of which may improve the system. Then, only those that are shown to perform better than others may be used to design real-life trials. This approach, introduces an evidence-based method to select interventions to test, rather than relying only on the opinion of those designing a trial. Therefore, future trials may save on costs and time if they focus on prioritizing interventions shown to be most promising from simulations.

Perspectives

The study on the burden of journal peer review answered many questions in a macroscopic level, however it also left many others open. One may study the system more in detail to reveal how the imbalanced distribution of reviewer effort was created. For instance, is it because the majority of scientists reject most invitations to review and only a few accept most of them or because those publishing the most are asked to review the most? Simulations on pubmed data may also show whether this distribution can be explained by the fact that most of researchers are unknown to the journal databases and to editors. Finally, one may test whether a central database of reviewers, which keeps track of the amount of reviews they perform, may facilitate the editorial task of finding the most appropriate reviewers without overburdening them.

Due to the flexibility and general nature of the agent-based model there are many interesting paths of research that one may follow in the future. First, the behavior of authors and reviewers during the peer-review process could be made more detailed. For instance, reviewers might follow personalized strategies; to cooperate with certain authors or to compete with others. Authors, on the other hand, may be tuned to reciprocate any behavior they encountered from past reviewers when they become reviewers themselves. Moreover, one could model the journals to be able to invite for peer review only those authors who exist in their databases (i.e, they have previously submitted to the journal). Invited reviewers may also be allowed to reject these invitations thus making it possible to model the struggle of editors to find referees.

Another way to increase complexity in the model and conduct a more detailed study of its variables, could be to split the resources and Q score variables into subcomponents. For instance, instead of resources, one may use metrics of funding, facilities and personnel, among other characteristics that aid the production of scientific results. Further, in lieu of Q score, one may use metrics of innovation, significance of results and level of bias among other characteristics that describe the intrinsic quality of a paper. In addition, papers may also have co-authors.

On the macroscopic scale, one could modify the model to measure additional interventions, such as a system in which authors submit their manuscripts to a central database and journals 'bid' to take them. Another idea would be to simulate different domains of science and measure their interactions. Finally, one may model a generalization of the agent-based model, in even larger scale, in which several alternative peer-review systems are jointly implemented. Then, scenarios that assume different adoption rates of alternative systems by journals may be introduced. This would allow the effect of the systems on overall peer-review efficiency, reviewer effort and scientific dissemination as a function of their adoption rate, to be studied.

The future of peer review and scientific publication

Currently, scientific publication looks very different from what it was twenty years ago. As the internet brought many changes to the system it is likely that other kind of technologies, such as artificial intelligence (AI), will affect the system importantly in the following years. Here, I will try to imagine how the system might look like in the near future.

First, I believe that open access is likely to completely dominate scientific publication. More and more journals are currently adopting this format and even funding agencies, such as Bill and Melinda Gates foundation and the European Union, require all their funded research to be open access (Bill and Melinda Gates foundation open access policy 2015; Enserink 2016). Moreover, review sharing and crowdsourcing of online reviews are going to be more widely adopted than they are now. These systems, seem to have many supporters, and technologically, they do not seem to pose a big challenge to the journals. Already more and more journals ask, at the time of submission, for permission to share any reviews with editors of other journals in case of rejection. Moreover, many open access journals and independent online services (such as PubPeer) currently allow for post-publication comments, which seems to be an increasing trend.

Second, artificial intelligence is expected to dominate our lives in the next few years and scientific publication will not be exempt from it. Difficult and time consuming editorial tasks such as finding reviewers will be automatized and optimized even more than they currently are. For example, editorial systems will be able to automatically analyze the content of a submitted manuscript and propose reviewers out of all literature and not limited to those existing in the journal's database. Even though AI may be expected to greatly help the editors, this is not exactly the case for the reviewers. Reviewing a paper is a much more difficult task for a computer than identifying reviewers is. However, AI can assist a lot the human reviewers by pointing out areas of a paper that might be problematic. For instance, a reviewer may be notified in cases such as when the word 'blinding' is missing from a paper reporting the results of a RCT.

Conclusions

In this PhD thesis, I studied the scientific publication and peer-review systems through a mathematical and complex-systems modeling approach. First, I focused on the burden that peer review has been posing to the scientific community. I showed that, contrary to many anecdotal claims, the scientific community can collectively meet the annual demand for peer review. However, only 20% of researchers have been performing 69%-94% of reviews.

Second, I developed an agent-based model (ABM) of the conventional scientific publication system calibrated with empirical data, pertaining to the biomedical domain. Third, using this ABM as a base, I developed five alternative peer-review systems and compared their performance in terms of peer-review efficiency, reviewer effort and scientific dissemination with that of the conventional system. My results indicated that each of the two review-sharing systems (cascade and portable peer review) showed important advantages, such as improved peer-review efficiency, less reviewer effort and better scientific dissemination. The rest of the systems, were either not as advantageous as the review-sharing ones (crowdsourcing and immediate publication) or performed worse than the conventional one in terms of peer-review efficiency (re-review opt-out). Therefore, I propose that the two review-sharing systems may be introduced into real-world trials.

The main advantage of my agent-based model is that its flexibility allows testing almost any kind of alternative system (or intervention), which has been proposed to improve peer review. Finally, even though scientific publication and peer-review systems are not perfect, this research shows that it is possible to improve them.

References

Alberts, B. (2013). Impact Factor Distortions. Science, 340(6134), 787-787, doi: http://dx.doi.org/10.1126/science.1240319.

Alberts, B., Hanson, B., & Kelner, K. L. (2008). Reviewing peer review. Science, 321(5885), 15, doi: http://dx.doi.org/10.1126/science.1162115.

Allesina, S. (2012). Modeling peer review: an agent-based approach. Ideas in Ecology and Evolution, 5(2), 27-35, doi: http://dx.doi.org/10.4033/iee.2012.5b.8.f.

Arns, M. (2014). Open access is tiring out peer reviewers. Nature, 515(7528), 467, doi: http://dx.doi.org/10.1038/515467a.

Auchincloss, A. H., & Diez Roux, A. V. (2008). A New Tool for Epidemiology: The Usefulness of Dynamic-Agent Models in Understanding Place Effects on Health. Am J Epidemiol, 168(1), 1-8, doi: http://dx.doi.org/10.1093/aje/kwn118.

Bill and Melinda Gates foundation open access policy (2015). http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy.

Björk, B. C., Roos, A., & Lauri, M. (2008). Global annual volume of peer reviewed scholarly articles and the share available via different Open Access options. In ELPUB2008. http://elpub.scix.net/cgi-bin/works/Show?178_elpub2008

Blank, R. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. The American Economic Review, 81(5), 1041-1067.

Bohannon, J. (2013). Who's Afraid of Peer Review? Science, 342(6154), 60-65, doi: `http://dx.doi.org/10.1126/science.342.6154.60`.

Bohannon, J. (2014). Google Scholar Wins Raves–But Can It Be Trusted? Science, 343(6166), 14, doi: `http://dx.doi.org/10.1126/science.343.6166.14`.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences, 99(suppl 3), 7280-7287, doi: `http://dx.doi.org/10.1073/pnas.082080899`.

Breuning, M., Backstrom, J., Brannon, J., Gross, B., & Widmeier, M. (2015). Reviewer Fatigue? Why Scholars Decline to Review their Peers' Work. PS: Political Science & Politics, 48(4), 595-600.

Bruce, R., Chauvin, A., Trinquart, L., Ravaud, P., & Boutron, I. (2016). Impact of interventions to improve the quality of peer review of biomedical journals: a systematic review and meta-analysis. BMC Medicine, 14(85), doi: `http://dx.doi.org/10.1186/s12916-016-0631-5`.

Callaway, E. (2015). Faked peer reviews prompt 64 retractions. Nature, doi: `http://dx.doi.org/10.1038/nature.2015.18202`.

Cals, J. W., Mallen, C. D., Glynn, L. G., & Kotz, D. (2013). Should authors submit previous peer-review reports when submitting research papers? Views of general medical journal editors. Ann Fam Med, 11(2), 179-181, doi: `http://dx.doi.org/10.1370/afm.1448`.

Cat, F., Adam, M., & Ivan, O. (2014). Publishing: The peer-review scam.

Nature, 515, 480–482, doi: http://dx.doi.org/10.1038/515480a.

Chhatwal, J., & He, T. (2015). Economic Evaluations with Agent-Based Modelling: An Introduction. PharmacoEconomics, 33(5), 423-433, doi: http://dx.doi.org/10.1007/s40273-015-0254-2.

Cohen, A., Pattanaik, S., Kumar, P., Bies, R. R., de Boer, A., Ferro, A., et al. (2016). Organised crime against the academic peer review system. Br J Clin Pharmacol, 81, 1012–1017, doi: http://dx.doi.org/10.1111/bcp.12992.

Csiszar, A. (2016). Peer review: Troubled from the start. Nature, 532, 306–308, doi: http://dx.doi.org/10.1038/532306a.

Culley, T. (2017). Want to peer review? Top 10 tips to get noticed by editors. Publons, https://publons.com/blog/10-things-you-need-to-know-to-get-noticed-by-editors.

Dallmeier-Tiessen, S., Darby, R., Hyppoelae, J., Igo-Kemenes, P., Kahn, D., Lambert, S., et al. (2010). First results of the SOAP project. Open access publishing in 2010. arXiv:1010.0506v1, https://arxiv.org/abs/1010.0506.

Davis, P. (2016). Are journals lacking for reviewers? The Scholarly Kitchen, https://scholarlykitchen.sspnet.org/2016/12/06/are-journals-lacking-for-reviewers/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+ScholarlyKitchen+%28The+Scholarly+Kitchen%29.

Day, T. E. (2015). The big consequences of small biases: A simulation of peer review. Research Policy, 44(6), 1266-1270, doi: http://dx.doi.org/10.1016/j.respol.2015.01.006.

Enserink, M. (2016). E.U. urged to free all scientific papers by 2020. Science, doi: http://dx.doi.org/10.1126/science.aaf9911.

Epstein, J. M. (2006). Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton, NJ: Princeton, NJ: Princeton University Press.

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. Nature, 460(7256), 685-686, doi: `http://dx.doi.org/10.1038/460685a`.

Garfield, E. (2006). The history and meaning of the journal impact factor. JAMA, 295(1), 90-93, doi:10.1001/jama.295.1.90.

group, T. F. (2016). Peer review: A global view. Taylor & Francis Group.

Gura, T. (2002). Scientific publishing: Peer review, unmasked. Nature, 416(6878), 258-260, doi: `http://dx.doi.org/10.1038/416258a`

Helmer, M., Schottdorf, M., Neef, A., & Battaglia, D. (2017). Gender bias in scholarly peer review. [JOUR]. eLife, 6, e21718, doi: `http://dx.doi.org/10.7554/eLife.21718`.

Herron, D. M. (2012). Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review. Surgical Endoscopy, 26(8), 2275-2280, doi: `http://dx.doi.org/10.1007/s00464-012-2171-1`.

Hopewell, S., Collins, G. S., Boutron, I., Yu, L.-M., Cook, J., Shanyinde, M., et al. (2014). Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. [Journal Article]. BMJ, 349, doi: `http://dx.doi.org/10.1136/bmj.g4145`.

Houry, D., Green, S., & Callaham, M. (2012). Does mentoring new peer reviewers improve review quality? A randomized trial. BMC Medical Education, 12(1), 83.

Hunter, J. (2012). Post-publication peer review: opening up scientific conversation. Frontiers in computational neuroscience, 6, 63.

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise. PLOS ONE, 11(11), e0166387, doi: http://dx.doi.org/10.1371/journal.pone.0166387.

Look, H., & Sparks, S. (2010). The value of UK HEIs contribution to the publishing process: Summary report. http://www.jisc-collections.ac.uk/Reports/valueofukhe/.

Marshall, B. D., & Galea, S. (2015). Formalizing the role of agent-based modeling in causal inference and epidemiology. Am J Epidemiol, 181(2), 92-99, doi: http://dx.doi.org/10.1093/aje/kwu274.

Martins, A. C. R. (2010). Modeling scientific agents for a better science. Advances in Complex Systems, 13(04), 519-533, doi: http://dx.doi.org/10.1142/s0219525910002694.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. Journal of the American Society for Information Science and Technology, 64(1), 132-161, doi: http://dx.doi.org/10.1002/asi.22798.

Nicolis, G., & Rouvas-Nicolis, C. (2007). Complex Systems. Scolarpedia, 2(11), 1473, doi: http://dx.doi.org/10.4249/scholarpedia.1473.

Oransky, I., & Marcus, A. (2016). Is the peer-review system sustainable? STAT, https://www.statnews.com/2016/11/18/peer-review-sustainable.

Osterath, B. (2016). Peer-review 'heroes' do lion's share of work. Nature News, doi: http://dx.doi.org/10.1038/nature.2016.21031.

Paolucci, M., & Grimaldo, F. (2014). Mechanism change in a simulation of peer review: from junk support to elitism. Scientometrics, 99(3), 663-688, doi: http://dx.doi.org/10.1007/s11192-014-1239-1.

Park, I.-U., Peacey, M. W., & Munafo, M. R. (2014). Modelling the effects of subjective and objective decision making in scientific peer review. [Letter]. Nature, 506(7486), 93-96, doi: http://dx.doi.org/10.1038/nature12786.

Peters, P., & Ceci, J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. Behavioral and Brain Sciences, 5(2), 187-195, doi: http://dx.doi.org/10.1017/S0140525X00011183.

Pontille, D., Torny, D. (2014). The blind shall see! The question of anonymity in journal peer review. Ada: A journal of Gender, New Media, and Technology, 4, doi: http://dx.doi.org/10.7264/N3542KVW.

Pöschl, U. (2012). Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. Frontiers in computational neuroscience, 6, 33.

Primack, R. B., Maron, M., & Campos-Arceiz, A. (2016). Who are our reviewers and how do they review? The profile and work of Biological Conservation reviewers. Biological Conservation, doi: http://dx.doi.org/10.1016/j.biocon.2016.11.039.

Rennie, D. (2003). Editorial peer review: its development and rationale. Peer review in health sciences, 2, 1-13.

Rennie, D. (2016). Let's make peer review scientific. [Comment]. Nature, 535(7610), 31–33, doi: http://dx.doi.org/10.1038/535031a.

Review rewards. (2014). [Editorial]. Nature, 514(274), doi: http://dx.doi.org/

10.1038/514274a.

Rigotti, N. A., & Wallace, R. B. (2015). Using Agent-Based Models to Address "Wicked Problems" Like Tobacco Use: A Report From the Institute of MedicineUsing Agent-Based Models to Address Tobacco Use. Ann Intern Med, 163(6), 469-471, doi: http://dx.doi.org/10.7326/m15-1567.

Roebber, P. J., & Schultz, D. M. (2011). Peer Review, Program Officers and Science Funding. PLOS ONE, 6(4), e18680, doi: http://dx.doi.org/10.1371/journal.pone.0018680.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., & Smith, R. (2004). Effects of training on quality of peer review: randomised controlled trial. BMJ, 328(7441), 673.

Sense About Science. (2004). Peer review and the acceptance of new scientific ideas: Sense About Science.

Sense About Science. (2012). Peer review: The nuts and bolts: Sense About Science. http://senseaboutscience.org/wp-content/uploads/2016/09/peer-review-the-nuts-and-bolts.pdf

Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. Journal of the Royal Society of Medicine, 99(4), 178-182.

Sorokowski, P., Emanuel, K., Agnieszka, S., & Katarzyna, P. (2017). Predatory journals recruit fake editor. Nature, 543, 481–483, doi: http://dx.doi.org/10.1038/543481a.

Spier, R. (2002). The history of the peer-review process. Trends in Biotechnology, 20(8), 357-358, doi: http://dx.doi.org/10.1016/S0167-7799(02)01985-6.

Squazzoni, F., & Gandelli, C. (2013). Opening the black box of peer review:

An agent-based model of scientist behaviour. Journal of Artificial Societies and Social Simulation, 16(2), 3.

Stahel, P. F., & Moore, E. E. (2014). Peer review for biomedical publications: we can improve the system. BMC Medicine, 12(179), doi: `http://dx.doi.org/10.1186/s12916-014-0179-1`.

Subert, P. (2012). Open access. The MIT Press Essential Knowledge Series.

Thurner, S., & Hanel, R. (2011). Peer-review in a world with rational scientists: Toward selection of the average. The European Physical Journal B, 84(4), 707-711, doi: `http://dx.doi.org/10.1140/epjb/e2011-20545-7`.

van Noorden, R. (2013). Company offers portable peer review. Nature, 494(7436), 161.

van Noorden, R. (2014). Publishers withdraw more than 120 gibberish papers. Nature, doi: `http://dx.doi.org/10.1038/nature.2014.14763`.

van Rooyen, S., Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers'recommendations: a randomised trial. [Journal Article]. BMJ, 318(7175), 23-27, doi: `http://dx.doi.org/10.1136/bmj.318.7175.23`.

Walker, R., & Rocha da Silva, P. (2015). Emerging trends in peer review: a survey. Frontiers Neurosci, 9, 169, doi: `http://dx.doi.org/10.3389/fnins.2015.00169`.

Ware, M., & Mabe, M. (2015). The STM Report: An overview of scientific and scholarly journal publishing (Fourth edition ed.). The Hague, The Netherlands: International Association of Scientific, Technical and Medical Publishers.

Warne, V. (2016). Rewarding reviewers – sense or sensibility? A Wiley study explained. Learn Publ, 29(1), 41-50, doi: http://dx.doi.org/10.1002/leap.1002.

SUPPLEMENTARY INFORMATION

S1 Appendix. Analytical methods

In this section, we first present how we estimated the number of submissions, the demand for peer review, the supply of peer review and the time researchers devote to peer review. Then we present the data we used to inform our modelling. Finally, we present sensitivity analyses over different distributions and values.

Estimation of demand and supply for peer-review

Let us consider N_p the number of articles accepted for publication. Let N_u be the number of articles submitted for publication but that ultimately remain unpublished. We accounted for multiple submissions after rejections, which all occurred within a given year. We assumed that both published and unpublished papers followed the same distribution of resubmissions. Let us define $R_i{}'$, the proportion of manuscripts submitted exactly *i* times. The proportion of manuscripts submitted at least i times is $R_i = \sum_{k\geq i} R'_{k-1}$. Then the total number of submissions is:

$$N_s = \left(N_p + N_u\right) \times \sum_{i=1}^{I} R_i \times i \qquad (1)$$

For simplicity, we set a maximum amount of resubmissions (*I*). For example, if 5% of papers are submitted once, 10% are submitted twice and 85% are submitted three times, then $R'_1 = 0.05$, $R'_2 = 0.10$, $R'_3 = 0.85$, $R_1 = 1$, $R_2 = 0.95$, and $R_3 = 0.85$. Then, $\sum_{\boldsymbol{i=1}}^{\boldsymbol{3}} \boldsymbol{R_i} \times \boldsymbol{i} = 1\times 1 + 0.95\times 2 + 0.85\times 3 = 5.45$. If we further assume that 800 manuscripts were ultimately published and 200 ultimately unpublished, the total number of submissions is $N_s = 800\times(1 + 0.95\times 2 + 0.85\times 3) + 200\times(1 + 0.95\times 2 + 0.85\times 3) = 1{,}000\times 5.45 = 5{,}450\ submissions.$

The distribution of resubmissions of published and unpublished papers might differ, but we can transform it to be the same:

$$N_u{}^0 \times \sum_{i=1}^{I} R_i^0 \times i = N_u{}^0 \times \alpha \times \sum_{i=1}^{I} R_i \times i = N_u \times \sum_{i=1}^{I} R_i \times i \qquad (2)$$

where α is a constant, $\mathrm{N_u}{}^0 = \frac{N_u}{\alpha}$ the real amount of unpublished papers and $\mathrm{R_i^0}$ the real proportion of papers (re)submitted i times but never published. For example, if $\mathrm{R_1^0} = 1$, $\mathrm{R_2^0} = 0.85$, and $\mathrm{R_3^0} = 0.55$, then $\sum_{\boldsymbol{i=1}}^{\boldsymbol{3}} \mathrm{R_i^0} \times \boldsymbol{i} = 4.35$. If $\mathrm{N_u}{}^0 = 100$, then the total number of submissions which did not result in a publication is 370. In reality we do not know both $\sum_{\mathrm{i=1}}^{\mathrm{I}} \mathrm{R_i^0} \times \mathrm{i}$ and $\mathrm{N_u}{}^0$ and it would be impossible to obtain reliable data for them. However, we know $\sum_{\mathrm{i=1}}^{\mathrm{I}} \mathrm{R_i} \times \mathrm{i}$ and we can represent $\sum_{\mathrm{i=1}}^{\mathrm{I}} \mathrm{R_i^0} \times \mathrm{i}$ in terms of it using a constant α. Then, we can group α and $\mathrm{N_u}{}^0$ into a single constant $\mathrm{N_u}$ and work with equation 1.

We estimated the annual demand for reviews $\mathrm{N_{reviews}}$ as:

$$N_{reviews} = (1-d) \times r_s \times (N_s + \sum_{i=1}^{I} S_i) \qquad (3)$$

where d is the proportion of desk-rejected submissions, r_s the number of reviewers per peer review round and S_i the amount of papers that went to a second round of peer review in their i^{th} (re)submission. We defined S_i as follows:

$$S_i = \beta \times (N_p + N_u) \times R_i \qquad (4)$$

where β is the probability of a second peer-review round per submission that is not desk-rejected.

We can estimate $N_{reviews}$ using a different formula, which this time involves the annual demand for reviewers $N_{reviewers}$.

$$N_{reviews} = N_{reviewers} \times \sum_{j=1}^{J} P_j \times j \qquad (5)$$

where J is the maximum amount of annual reviews that any reviewer performed, j the amount of reviews completed from a reviewer in a given year and P_j the proportion of reviewers who completed j reviews. For example, if 1,000 scientists reviewed at least one paper inside a year, 60% of them performed 1 and 40% of them 2 reviews, then $\mathrm{N_{reviews}} = 1000 \times (0.6 \times 1 + 0.4 \times 2) = 1{,}400\ reviews$. Since we have two formulas estimating $N_{reviews}$, we can estimate the annual demand for reviewers from their combination:

$$N_{reviewers} = \frac{N_{reviews}}{\sum_{j=1}^{J} P_j \times j} = \frac{(1-d) \times r_s \times (N_s + \sum_{i=1}^{I} S_i)}{\sum_{j=1}^{J} P_j \times j} \qquad (6)$$

We defined each researcher's total amount of time available for research as follows:

$$t_{res} = work\ time \times (year - weekends - holidays) \qquad (7)$$

Collection and use of data

All data and results can be found in the accompanying Excel file (http://www.clinicalepidemio.fr/peerreview_burden/). We programmed our simulations by using MATLAB (MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Natick, MA, USA). The code is available at https://github.com/kovanostra/global-burden-of-peer-review.

We used data pertaining to the biomedical domain, except to estimate r_s and the distribution of peer-review effort ($\sum_{j=1}^{J} P_j$), for which we used data pertaining to all scientific disciplines. We extracted all records indexed as "journal articles" by MEDLINE from January 1, 1990 to December 31, 2015. We downloaded the xml files for each year separately and parsed them by using a script written in Python (also available on github). We excluded all records with no author name (*e.g.,* less than 0.001% of all articles for 2015) and indexed all authors based on their "LastName", "ForeName" and "Initials". We counted all the unique occurrences of authors by taking into account all these three pieces of information. For missing "ForeName" and/or "Initials", we used only the available fields. We did not use any methods for author name disambiguation for researchers indexed under the same "LastName", "ForeName" and "Initials". [13, 14] We set N_S to be equal to the number of publications for which we identified at least one author.

We assumed that potential reviewers in a given year were researchers who co-authored at least one paper that year (Scenario 1). Then we defined more stringent scenarios (in terms of which co-authors are potential reviewers) whereby candidate reviewers were the first or last authors of any article during the previous 3 years (Scenario 2); the first, second or last authors for the same year (Scenario 3); and the first or last authors for the same year (Scenario 4). For Scenario 2, we arbitrarily chose a time window of 3 years, which however may reflect changes in the databases that editors use to find reviewers. For each scenario, we repeated the same procedure of identifying the unique occurrences of authors as described above. For each scenario, the number of authors obtained was considered to represent the potential supply of reviewers ($N_{reviewers-supply}$) in any given year. We did not account for individual interactions between authors, editors and reviewers which may influence the potential supply of reviewers. We estimated the potential supply of reviews by using the relation $N_{reviews-supply} = N_{reviewers-supply} \times \sum_{j=1}^{J} P_j \times j$.

We obtained $\sum_{i=1}^{I} R_i$ and the empirical distribution of the time taken to perform each review from the 2009 Peer Review Survey, an international survey of 4,037 researchers [8]. Data corresponded to the biomedical domain. We considered r_S to be equal to 2.5 reviewers per peer-review round [11]. We obtained the empirical distribution of individual contributions to the peer-review effort ($\sum_{j=1}^{J} P_j$) for 2015 from the Publons reviewer recognition platform. In Publons, reviewers mainly self-report the reviews they have completed (ie, by forwarding review receipts to them). Publons was launched in 2012 and thus we could not obtain data for all unique years of our analysis. We assumed that the distribution for 2015 was identical for every year from 1990 to 2015. To our best knowledge, reliable data pertaining to β, N_u and d do not exist. We assumed that 90% of the peer-reviewed submissions went through a second round of peer review ($\beta = 0.9$), the percentage of the finally unpublished papers was equal to the 20% of the total submissions ($N_u = \gamma T_S$, $\gamma = 0.20$) and that the average proportion of papers desk-rejected was 25% ($d = 0.25$). Table 1 presents the values of the previously mentioned parameters.

Table 1: Parameter values

Variable	Description	Value	Source
r_s	Average reviewers per paper	2.5	Reference 4
β	Chance of second peer-review round	90%	No reference – Sensitivity analyses performed
γ	Proportion of unpublished papers among all submissions	20%	No reference – Sensitivity analyses performed
d	Average proportion of desk-rejected papers	25%	No reference – Sensitivity analyses performed
holidays	Holidays	25.3	Reference 5

For each researcher, we estimated the total amount of time available for research t_{res}, taking into account whether the researcher was full or part time. We used empirical data provided by the National Institute of Health and Medical Research of France (INSERM), which pertains to all its researchers. The total time spent in peer review was estimated by sampling the respective empirical distribution over the amount of reviews (j) completed by each reviewer. For example, if 65% of reviews required 1 to 5 hours to complete, 22% of them 6 to 10 etc., then for each review that a reviewer performed we first drew at random the duration range: between 1 and 5 hours with probability 65%, between 6 and 10 with probability 22%, etc. Afterwards, the actual review time was drawn from a uniform distribution over the interval. Comparing the time devoted to peer review with the total time available for research, we derived the proportion of researchers who devoted certain proportions of their time to peer review (full time, 50% or 30% of their annual work-time). For full-time workers, we used $work\ time = 8\ hours/day, year = 365\ days$ and $weekends = 104\ days$. We derived the amount of holidays by averaging between 21 OECD countries ($holidays = 25.3\ days$) [15]. For each full-time employed researcher, we obtained $t_{res} = 1{,}885\ hours$ and for part-time researchers $t_{res} = 943\ hours$ and $t_{res} = 566\ hours$ for those devoting 50% and 30% of their time to research, respectively.

S2 Appendix. Sensitivity analyses

We performed 25 sensitivity analyses in addition to our main analysis (Table 1). In the first 3 analyses, we used distributions of peer-review effort other than Publons 2015. Under the same conditions, we obtained the respective distributions from Publons for the years 2013 and 2014, corresponding again to all scientific domains. We also used a review effort distribution from only a single journal (*Nature Materials* 2002-2012).[6] Publons data concerned in total about 70,000 researchers and more than 10,000 journals, while data from Nature materials concerned about 4,500 and a single journal. For the remaining 22 sensitivity analyses, we varied the values of the parameters (*β, γ, d*) while using only the distribution from Publons 2015.

We evaluated all our sensitivity analyses under one outcome, the surplus in the annual number of potential reviewers as compared with the annual demand. For sensitivity analysis 1, we explored the possible surplus in the number of potential reviewers for each of the four scenarios as compared to the respective demand. For sensitivity analyses 2 to 25, we defined the surplus by using only scenario 4 and in some cases scenario 3 as well.

The distribution from *Nature Materials*, when using scenarios 1 and 2, produced a surplus for any given year (scenario 2 after 1999). However, it produced a deficit when considering scenarios 3 and 4 for any given year (Figure 1.A). When using scenario 4 the distributions of the peer-review effort from Publons for 2014 and 2013 produced surplus in the potential supply of reviews and reviewers when compared to scenario 3, for any given year, and when compared to scenario they produced surplus after 2001 and 2011, respectively, (Figure 1.B).

For most of the values of *γ*, we found a surplus in the number of available reviewers as compared to scenario 4, and for all of them when compared to scenario 3 (Figure 3 & 4). Variations over the values of *β* and *d* did not produce any deficit when compared to scenario 4 (except for d = 0.20 and before 2000) (Figure 2, Figure 5). Almost all sensitivity analyses (apart from the one of Nature Materials) for the last 3 years produced a surplus in the number of available reviewers, even though we compared them to the smallest pool of potential peer reviewers. Those that produced deficit when compared to scenario 4, always produced surplus when compared to scenario 3.

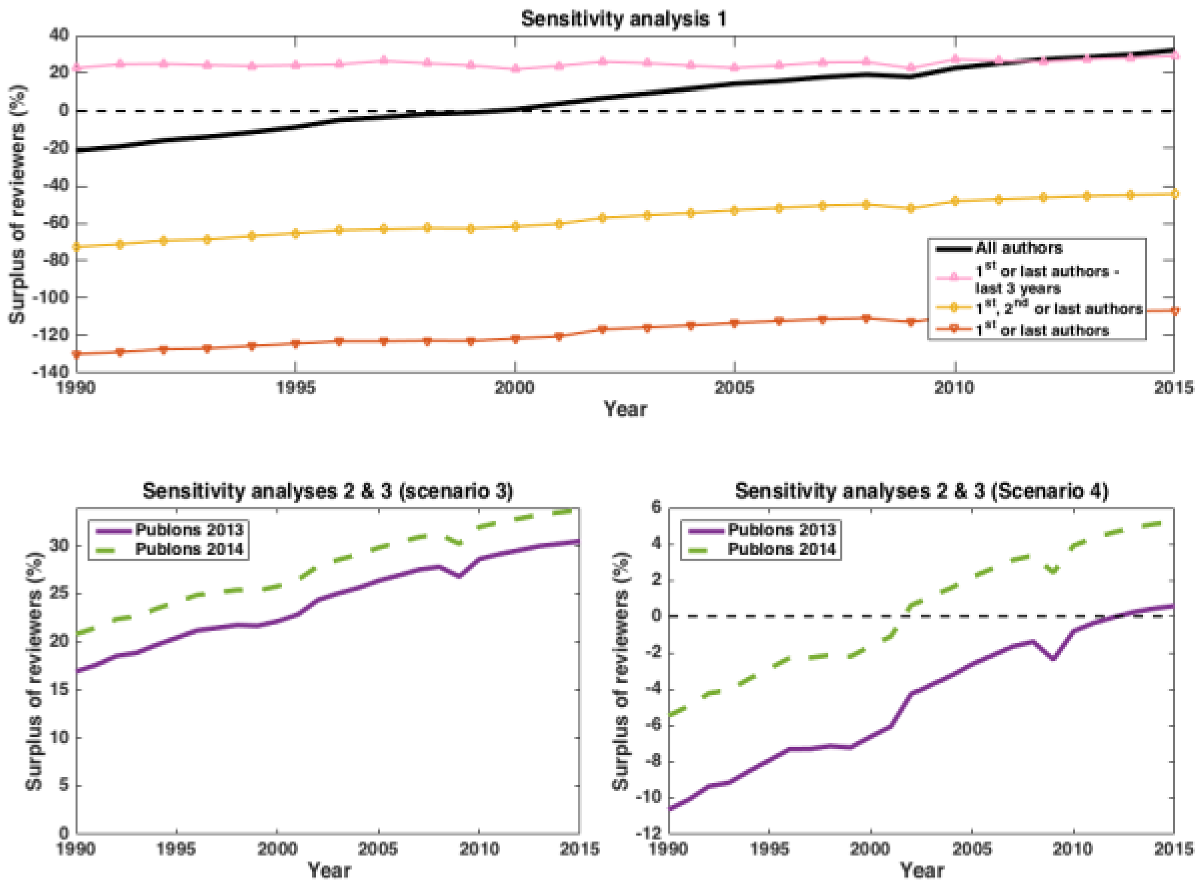
Table 2: Sensitivity analyses

Sensitivity analysis	Peer-review effort distribution	Desk rejection proportion *(d)*	Proportion of unpublished submissions to all submissions *(γ)*	Probability of second round of reviews *(β)*
Main analysis	Publons 2015	0.25	0.200	0.90
1	Nature Materials	″	″	″
2	Publons 2013	″	″	″
3	Publons 2014	″	″	″
4	Publons 2015	0.20	″	″
5	″	0.30	″	″
6	″	0.35	″	″
7	″	0.40	″	″
8	″	0.45	″	″
9	″	0.50	″	″
10	″	0.55	″	″
11	″	0.60	″	″
12	″	0.25	0.100	″
13	″	″	0.135	″
14	″	″	0.170	″
15	″	″	0.235	″
16	″	″	0.270	″
17	″	″	0.300	″
18	″	″	0.335	″
19	″	″	0.370	″
20	″	″	0.200	0.60
21	″	″	″	0.65
22	″	″	″	0.70
23	″	″	″	0.75
24	″	″	″	0.80
25	″	″	″	0.85

Distribution of peer-review effort and values for *d, γ, β* from the main analysis and from all sensitivity analyses.

Sensitivity analyses 1–3

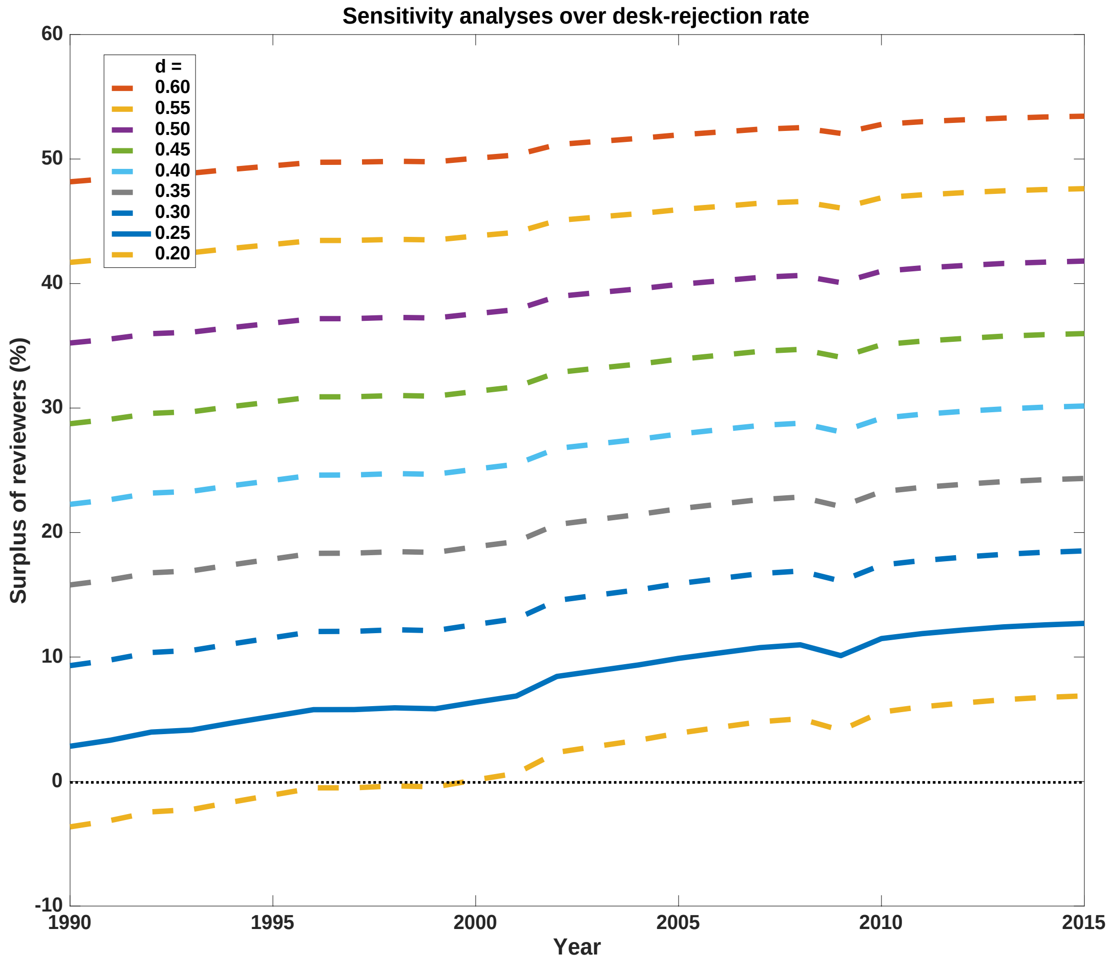
Figure 1. Sensitivity analyses by distributions of peer-review effort



A. Sensitivity analysis involved the distribution from *Nature Materials* (2002–2012). Surplus defined with all scenarios 1–4 to identify the potential supply of reviewers. **B.** Sensitivity analyses involved the distributions from Publons for the years 2013 and 2014. Surplus defined with only scenarios 3 and 4 to identify the potential supply of reviewers.

Sensitivity analyses 4–11

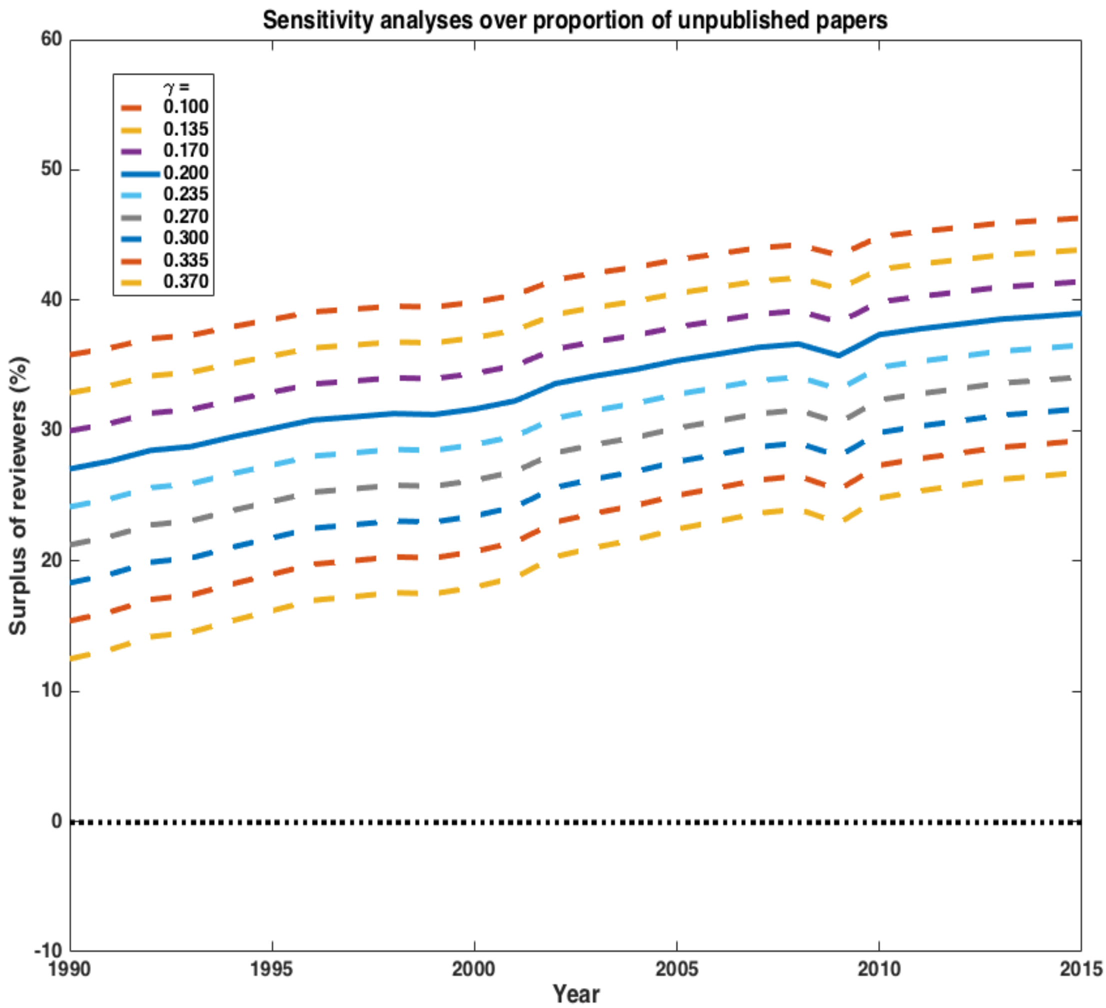
Figure 2. Sensitivity analyses by desk-rejection rate



Sensitivity analyses of different values of the overall proportion of desk-rejected manuscripts per submission. The continuous line shows the value used in the main analysis.

Sensitivity analyses 12–19

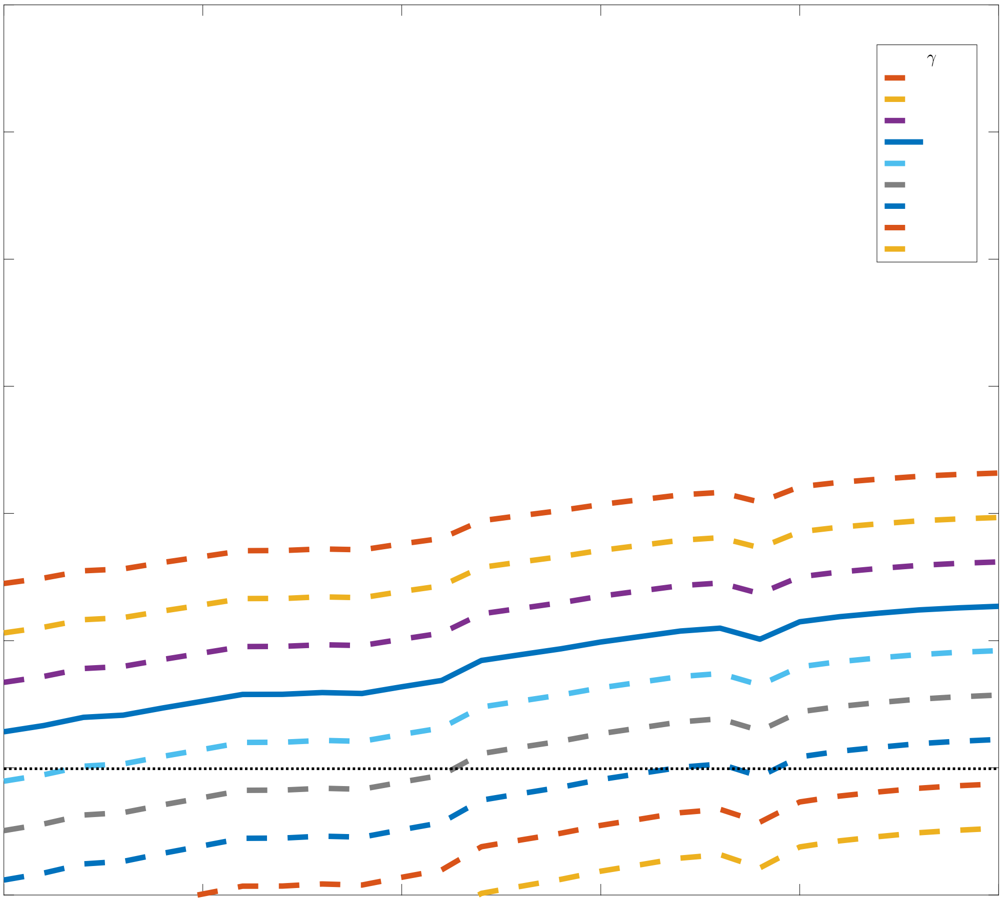
Figure 3. Sensitivity analyses by proportion of unpublished papers (Scenario 3)



Sensitivity analyses of different values of the proportion of unpublished papers compared to overall submissions. Surplus defined with scenario 3 to identify the potential supply of reviewers. The continuous line shows the value used in the main analysis.

Sensitivity analyses 12–19

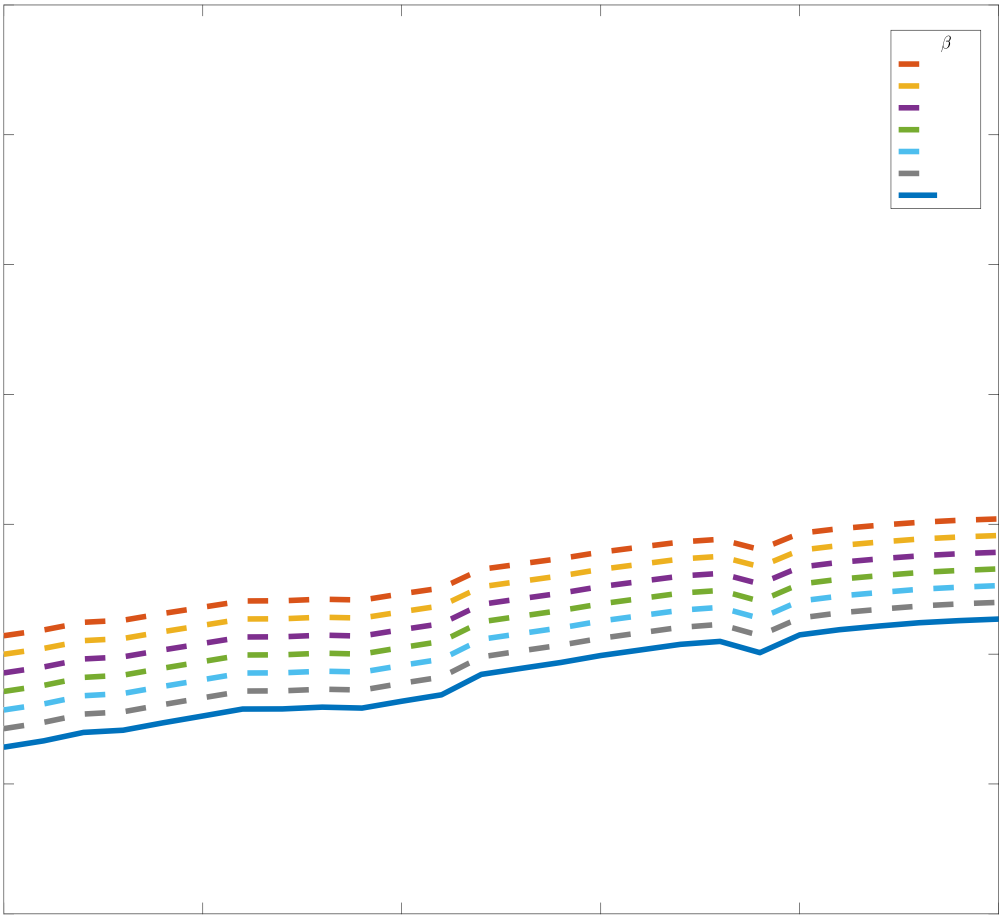
Figure 4. Sensitivity analyses by proportion of unpublished papers (Scenario 4)



Sensitivity analyses of different values of the proportion of unpublished papers compared to overall submissions. The continuous line shows the value used in the main analysis.

Sensitivity analyses 20–25

Figure 5. Sensitivity analyses by probability of second round of reviews



Sensitivity analyses of different values of the probability of papers going through a second round of peer review for a submission that was not desk-rejected. The continuous line shows the value used in the main analysis.

References

1. Lerchenmueller MJ, Sorenson O. Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists. PLoS ONE. 2016;11(7):e0158731.

2. Schulz C, Mazloumian A, Petersen AM, Penner O, Helbing D. Exploiting citation networks for large-scale author name disambiguation. EPJ Data Science. 2014;3(1):1-14. doi: 10.1140/epjds/s13688-014-0011-3.

3. Mulligan A, Hall L, Raphael E. Peer review in a changing world: An international study measuring the attitudes of researchers. J Assoc Inf Sci Technol. 2013;64(1):132-61. doi: 10.1002/asi.22798.

4. Look H, Sparks S. The value of UK HEIs contribution to the publishing process: Summary report: Rightscom Ltd for JISC Collections; 2010. Available from: http://www.jisc-collections.ac.uk/Reports/valueofukhe/.

5. Ray R, Schmitt J. No-vacation nation USA – a comparison of leave and holiday in OECD countries. EEE Policy Brief. 2007;3.

6. A decade in numbers. Nat Mater. 2012;11(9):743-4.

List of Figures