RareGems: A Proposed Database of Rare Disease Science and Innovation Landscape

Kevin Chandra[1]

Department of Strategy and Innovation

Questrom School of Business, Boston University

Since the passage of the Orphan Drugs Act of 1983, there has been a flurry of pharmaceutical innovation, specifically for rare diseases. This is illustrated by a considerable growth in the number of designations—and subsequent authorizations by the US Food and Drug Administration (FDA)—for drugs intended to serve these patient populations. As of July 2023, FDA has granted 6,756 orphan designations to 4,872 drug candidates, with 1,151 such designations eventually leading to full authorization for 715 drug candidates. Moreover, in recent years, nearly 50% of novel drugs approved by the agency have been those intended for rare disease populations (FDA, 2024).

Nevertheless, despite the growing interest among pharmaceutical and biotechnology firms, treatment options remain unavailable for a vast majority of rare diseases. Existing studies estimate that the figure could be as high as 90-95% (e.g. Wood et al., 2013; Kaufmann, Pariser, and Austin, 2018; National Organization for Rare Disorders, 2025). Moreover, rare diseases, in aggregate, are not as 'rare' as initially assumed. Depending on the utilized classifications (e.g. NIH GARD, Orphanet, Monarch Disease Ontology / MONDO), the number of rare diseases and related conditions ranges from 6,000 to over 10,000 (Richter et al., 2015; Haendel et al., 2020; Nguengang-Wakap et al., 2020). Overall, these diseases are estimated to affect a total of 300-450 million people globally (Nguengang-Wakap et al., 2020), and about 30 million in the United States alone (National Organization for Rare Disorders, 2025). In other words, there remains a vast unmet need in terms of addressing health challenges faced by these patient populations.

The contrast between the growing number of orphan drug-related innovations and the persistent unmet need among most rare disease patients implies the need for a systematic documentation of the rate and direction of science and innovation in this space. However, to the best of my knowledge, there have been no studies that provide such contribution. Previous studies are predominantly focused on highly prevalent diseases, i.e. health conditions with large patient populations (e.g. Cockburn and Henderson, 1998; Acemoglu and Linn, 2004). Works that explore pharmaceutical innovation in rare diseases largely emphasize either the role of policy incentives, such as the role of Orphan Drugs Act in incentivizing orphan drug applications (Yin, 2008) and the impacts of FDA Breakthrough Therapy Designation on drug approvals (Chandra et al., 2024), or the challenges in developing and disseminating such treatments, e.g. Olsder, Martagan, and Tang (2023) on how pricing dynamics and patient populations influence firms' (dis)incentives, or Garassino et al. (2025) on how the current drug approval system renders the development of individualized therapies economically infeasible. While these studies have improved our understanding of the role of policy incentives and challenges entailed in incentivizing such innovations, they are primarily focused on successful, eventually approved drugs that cover a limited proportion of rare disease patient populations.

From another angle, other studies also pay attention to the role of patient advocacy organizations in shaping the allocation of resources to support disease-specific research, e.g. Best (2012) and

[1] The author received no financial support for the research and authorship of this extended abstract.

Hegde and Sampat (2015) on the influence of advocacy groups on NIH funding. While these studies have demonstrated a positive correlation, it remains to be seen whether an infusion of additional funding, in itself, is sufficient to incentivize scientists to enter these fields, let alone to pioneer scientific breakthroughs with high commercialization potential. Moreover, even if scientists decide to redirect their focus, the switching costs could be high (Myers, 2020), especially due to inherent uncertainties in studying rare diseases. In addition, there has been a growing interest among economists and policymakers in estimating the returns of government-funded research investments, and social welfare more broadly (e.g. Bryan and Williams, 2021; Jones and Summers, 2022).

To address these gaps, I propose to develop "RareGems", a database that highlights the science and innovation landscape for rare diseases. Combining the approaches of Azoulay et al. (2019), Bikard and Marx (2020), and Ewens and Marx (2024), I plan to develop a systematic recording of linkages between NIH grants, publications, patents that cite these papers, the formation of biotech firms catering towards these diseases, clinical trials, and subsequent drug applications and/or authorizations from the FDA. As patient advocacy organizations also play a major role in driving public attention to these diseases (Best, 2012; Wood et al., 2013; Hegde and Sampat, 2015), I also propose to examine their involvement in various aspects, e.g. providing seed funding for scientists, sponsoring clinical trials for patients, hiring scientists and/or industry actors into their advisory boards, or investing in early-stage biotechnology firms (as demonstrated by the venture philanthropy model adopted by Cystic Fibrosis Foundation, with Vertex Pharmaceuticals among their major success stories).

At the current stage, I am collecting data from various sources in order to construct this database. For the disease list, I currently rely on the National Organization for Rare Disorders (NORD) In-Depth Reports, which provides detailed elaboration on a total of 1,411 rare diseases as of this year. In addition, the In-Depth Reports also provide a list of equivalent medical keywords for each of these diseases. While these reports only cover a small fraction of all rare diseases identified in other classifications, e.g. NIH GARD, Orphanet, or MONDO, they have been used in various empirical studies and are often referred to as a 'go-to' source for patients and their family members (Yin, 2008; Institute of Medicine of the National Academies, 2010; Hegde and Sampat, 2015).

Afterwards, I utilize these disease names and their equivalent keywords to collect the relevant publication and patent data. I am currently collecting these data from the Lens.org database. The database has been especially important for public health and biomedical researchers as it amalgamates MeSH (Medical Subject Headings) key terms, Microsoft Academic Graph (MAG) fields of study, and keywords automatically generated by the Lens.org algorithm. This is demonstrated by several recent works, e.g. mapping the innovation landscape of CRISPR gene-editing tools (Jefferson et al., 2021), or analyzing the use of chemical substances for drug development (Wong and Yeong, 2021). Specific to the patent data, I would also focus on collecting: a) patents that cite these disease-specific publications, regardless of their relevance to the focal disease, and; b) patents that contain these disease-specific keywords, regardless of their citation linkages to these focal-disease publications. The resulting rare disease-specific patents, therefore, would be those that fulfill both criteria (citing disease-specific papers and containing at least one keyword of the same focal disease). While the focus of data collection is global, I will pay particular attention to USPTO patent applications that are subsequently granted, and more broadly, research that involves at least one researcher based in the United States.

Once the collection process is completed, I plan to link these patent-to-paper citations to the Reliance on Science dataset (Marx and Fuegi, 2020; Marx and Fuegi, 2022), which contains information on: a) applicant/examiner flags; b) whether citations are added to the front page, body text, or both; c) journal impact factor (possibly a proxy for the quality of these papers), and; d) journal commercial impact factor (a proxy for their potential commercializability). While their data are linked to paper IDs from OpenAlex (successor to the Microsoft Academic Graph), I plan to cross-walk the linkages using the papers' DOI identifiers, which are available in both the Reliance on Science and Lens.org. Moreover, I would also link these USPTO patent grants to the assignees in the Ewens and Marx (2024) dataset, with a particular interest in identifying VC-funded firms that are developing treatments and/or other medical devices to address these diseases. Additionally, in measuring the economic value of these patents, I would also link these patents to the KPSS dataset (Kogan et al., 2017), the latest version of which covers patents granted by the end of 2023.

Following Azoulay et al. (2019), I also plan to link these patent-paper pairs to the NIH grants that result in these publications. The primary source for these data is NIH RePORTER, which contains detailed information on NIH-funded projects from 1985 onward. Utilizing the same disease names and their equivalent medical keywords from NORD, I am currently extracting data on these disease-specific projects. The information of particular interest is the total cost (in nominal US$), which, once aggregated at the disease level, is of considerable importance in estimating the economic impact of public research funding for each of these diseases.

With regard to the advocacy organization data, I am collecting the list of relevant entities from the NORD website. At present, their database provides information on over 2,000 patient organizations, hospitals, as well as government agencies, with over 350 of them registered as NORD's official organization members. I would then match these organization names to the compiled data in various aspects, such as funding acknowledgements of these papers (as collected from Lens.org), sponsorship in clinical trials (with the data to be collected from clinicaltrials.gov), and linkages to other institutions by examining the presence of scientists and/or industry representatives in either the boards of directors or scientific/medical advisory boards of these entities.

In the near future, I plan to complete the RareGems database construction by linking these integrated data to clinical trials and drug development projects. The data for the former will be crucial in understanding the nature of inter-organizational collaboration between advocacy groups, universities, and biotech/pharma firms, and the ways in which such collaboration influences the subsequent clinical trial outcomes. Specific to the latter, my emphasis will be on applications for orphan drug designations, as well as those that eventually lead to FDA's full authorization. The purpose is four-fold: a) to identify companies that are fully engaged from upstream (basic research, as attested by their scientific publications) up to commercialization; b) to understand the competitive landscape among these firms across disease-specific research ecosystems; c) to better capture the direct impacts of NIH funding and/or support provided by advocacy organizations on subsequent drug development, and; d) to eventually link these drug names to various other economic indicators beyond KPSS patent value, such as annual sales, Medicare/Medicaid claims, or the long-term survival of these patients.

I foresee that RareGems will be highly relevant for academics, industry actors, as well as patient communities. Specific to academics, the database will be of particular relevance for innovation and health economists interested in examining the direct impacts and spillovers of health research to other fields, as well as their estimated economic value. RareGems can also serve as a useful

database for strategy and management scholars interested in the landscape of competition among pharma/biotech firms, and inter-organizational relations more broadly (especially with advocacy groups and policymakers). Industry actors, and especially those involved in drug development, will find this database appealing as they are able to better understand the private value of public research, as well as the landscape of scientific research and innovative activities among their competitors. RareGems will be of special importance for patient communities, as the database provides information about the current state of scientific literature, and better inform their diagnostic processes, which are frequently time-consuming and prone to misdiagnoses (Wood et al., 2013). Rare disease patients, in particular, highly rely on scientific publications due to diffculties in receiving accurate diagnoses, in general, from physicians (Wood et al., 2013).

With regard to dissemination, my tentative plan is to make the data public either in 2026 or 2027. In addition. I plan to reach out to the team at Lens.org, and to work closely with them, in order to facilitate the creation of a special, constantly-updated Lens Report. This report describes the landscapes of rare disease research and innovation in a way similar to their 'Landscapes of Climate Innovations' (https://www.lens.org/lens/landscape). I would also reach out and work with the team at NORD for facilitating outreach to the broader patient communities in accessing the RareGems data, which I propose will be hosted under the Lens.org website.

REFERENCES

Acemoglu, D., and Linn, J. (2004). Market size in innovation: Theory and evidence from the pharmaceutical industry. *Quarterly Journal of Economics*, 119(3): 1049-1090.

Azoulay, P., Graff-Zivin, J. S., Li, D., and Sampat, B. N. (2019). Public R&D investments and private-sector patenting: Evidence from NIH funding rules. *Review of Economic Studies*, 86(1): 117-152.

Best, R. K. (2012). Disease politics and medical research funding: Three ways advocacy shapes policy. *American Sociological Review*, 77(5): 780-803.

Bikard, M., and Marx, M. (2020). Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, 66(8): 3425-3443.

Bryan, K. A., and Williams, H. L. (2021). Innovation: market failures and public policies. In K. Ho et al. (eds), *Handbook of industrial organization*, pp. 281-388. Amsterdam: North-Holland.

Chandra, A., Kao, J., Miller, K. L., and Stern, A. D. (2024). Regulatory incentives for innovation: the FDA's Breakthrough Therapy Designation. *The Review of Economics and Statistics*, DOI: 10.1162/rest_a_01434.

Cockburn, I. M., and Henderson, R. M. (1998). Absorptive capacity, co-authoring behavior, and the organization of research in drug discovery. *The Journal of Industrial Economics*, 46(2): 157-182.

Ewens, M., and Marx, M. (2024). Firm age and invention: An open-access dataset. Working paper, link retrieved from https://foundingpatents.com/

FDA (Food and Drug Administration) (January 9, 2024). 'FDA Approves Many New Drugs in 2023 that Will Benefit Patients and Consumers'. Link retrieved from https://www.fda.gov/news-events/fda-voices/fda-approves-many-new-drugs-2023-will-benefit-patients-and-consumers

Garassino, M. C., Odunsi, K., Siniscalchi, M., and Veronesi, P. (2025). On the economic infeasibility of personalized medicine, and a solution. NBER Working Paper 33539, link retrieved from https://www.nber.org/papers/w33539

Haendel, M., Vasilevsky, N., Unni, D., Bologa, C., Harris, N., Rehm, H., Hamosh, A., Baynam, G., Groza, T., McMurry, J., Dawkins, H., Rath, A., Thaxon, C., Bocci, G., Joachimiak, M. P., Kohler, S., Robinson, P. N., Mungall, C., and Oprea, T. I. (2020). How many rare diseases are there? *Nature Reviews Drug Discovery*, 19(2): 77-78.

Hegde, D., and Sampat, B. N. (2015). Can private money buy public science? Disease group lobbying and federal funding for biomedical research. *Management Science*, 61(10): 2281-2298.

Institute of Medicine of the National Academies (2010). *Rare diseases and orphan products: Accelerating research and development*. Washington, D.C.: National Academies Press.

Jefferson, O. A., Lang, S., Williams, K., Koellhofer, D., Ballagh, A., Warren, B., Schellberg, B., Sharma, R., and Jefferson, R. A. (2021). Mapping CRISPR-Cas9 public and commercial innovation using The Lens institutional toolkit. *Transgenic Research*, 30(4): 585-599.

Jones, B. F., and Summers, L. H. (2022). A calculation of the social returns to innovation. In A. Goolsbee and B. F. Jones (eds.), *Innovation and public policy*. Chicago, IL: University of Chicago Press.

Kaufmann, P., Pariser, A. R., and Austin, C. (2018). From scientific discovery to treatments for rare diseases—the view from the National Center for Advancing Translational Sciences: Office of Rare Diseases Research. *Orphanet Journal of Rare Diseases*, 13(196), DOI: 10.1186/s13023-018-0936-x.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, 132(2): 665-712.

Marx, M., and Fuegi, A. (2020). Reliance on science: Worldwide front-page citations to scientific articles. *Strategic Management Journal*, 41(9): 1572-1594.

Marx, M., and Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2): 369-392.

Myers, K. R. (2020). The elasticity of science. *American Economic Journal: Applied Economics*, 12(4): 103-134.

Nguengang-Wakap, S., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., Murphy, D., Le Cam, Y., and Rath, A. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics*, 28(2): 165-173.

Olsder, W., Martagan, T., and Tang, C. S. (2023). Improving access to rare disease treatments: Subsidy, pricing, and payment schemes. *Management Science*, 69(9): 5256-5274.

Richter, T., Nestler-Parr, S., Babela, R., Khan, Z. M., Tesoro, T., Molsen, E., and Hughes, D. A. (2015). Rare disease terminology and definitions—A systematic global review: report of the ISPOR Rare Disease Special Interest Group. *Value in Health*, 18: 906-914.

Wong, X. K., and Yeong, K. Y. (2021). A patent review on the current developments of benzoxazole. *ChemMedChem*, 16(21): 3237-3262.

Wood, J., Sames, L., Moore, A., and Ekins, S. (2013). Multifaceted roles of ultra-rare and rare disease patients/parents in drug discovery. *Drug Discovery Today*, 18(21-22): 1043-1051.

Yin, W. (2008). Market incentives and pharmaceutical innovation. *Journal of Health Economics*, 27: 1060-1077.