A linked data approach to investigate the history of rare diseases and the discovery of their genetic cause

Friederike Ehrhart[1,2], Egon L. Willighagen[1], Nasim Bahram Sangani[1,2], Martina Kutmon[1,3], Leopold M.G. Curfs[2] and Chris T. Evelo[1,2,3]

1 Dept. of Bioinformatics, NUTRIM, Maastricht University, Maastricht, The Netherlands; 2 GKC-Rett Expertise Centre, Maastricht University Medical Centre, Maastricht, The Netherlands; 3 Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, The Netherlands

Background: Rare diseases have been observed and documented since the ancient. Nevertheless, only since the development of molecular biology methods in the last century it was possible to identify and investigate their underlying genetic causes. In this study we collected and investigated first documentations of rare diseases and the discovery of their genetic cause and used this information for further analysis.

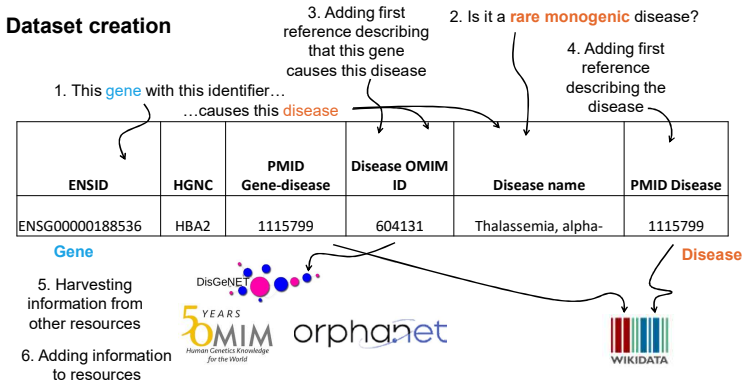

THE BRITISH MEDICAL JOURNAL [June 30, 1888.

ON RARE DISEASES AND EXCEPTIONAL SYMPTOMS.

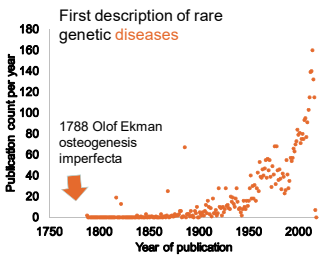
By JONATHAN HUTCHINSON, F.R.C.S., F.R.S., LL.D., Emeritus Professor of Surgery at the London Hospital.

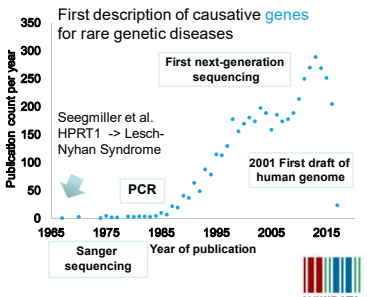
Methods: Data and information about rare genetic diseases, their causative genes and literature information about the first publication were collected from OMIM, Whonamedit, PubMed, and Google scholar. The dataset was constructed and harmonised in a spreadsheet and as machine-readable RDF nanopublication. The data is available in a Figshare data collection. The acquired data identifiers were then used to harvest information from other resources like Wikidata, DisGeNET, and Orphanet.

Dataset creation



Results: According to this underlying data, the description of rare genetic diseases started in 1788 with osteogenesis imperfecta. The first discovery of a causative gene was in 1967 with the gene causing Lesch-Nyhan syndrome. Investigating the timeline, the discovery rate of genes is linked to developments in molecular biology techniques while first descriptions of rare diseases follow the general trends in publication numbers.

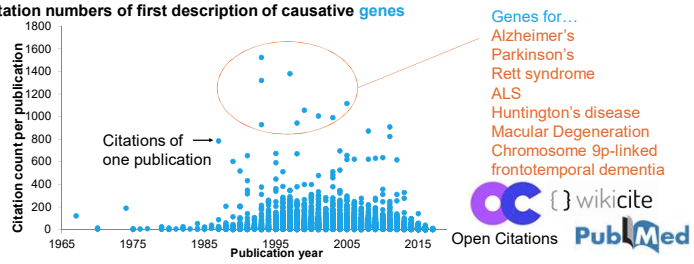




Where are new rare disease or a causative gene published?

Journal	Count	Journal	Count
First description of a new disease (100% = 3144)		First description of new genes causing a rare disease (100% = 4263)	
American Journal of Human Genetics	457	American Journal of Human Genetics	1137
Nature Genetics	145	Nature Genetics	786
American Journal of Medical Genetics Part A	137	Human Molecular Genetics	266
Journal of Medical Genetics	136	Journal of Medical Genetics	175
Human Molecular Genetics	122	The New England Journal of Medicine	148
The New England Journal of Medicine	121	Journal of Clinical Investigation	136
Journal of Clinical Investigation	78	Proceedings of the National Academy of Sciences of the United States of America	129
Neurology	77	Science	121
The Journal of Pediatrics	68	Nature	99
Proceedings of the National Academy of Sciences of the United States of America	59	Human Mutation	94

Rare or truly neglected diseases? Citation scores shows…



Using identifier mapping, made available by DisGeNET, further information like disease prevalence data from ORPHANET, preferred publication journals from Wikidata, and disease super classes from DisGeNET could be acquired.

1. Identifier & entity mapping – many unique identifier systems for diseases, phenotypes, syndromes etc.
2. Harvesting interesting information

Disease OMIM ID	CUI	Disease SemanticType	MeSH subclass	ORPHA	Epidemiology
604131	C0002312	Disease or Syndrome	Nervous System Diseases [mesh:C10]	846	>1 / 1000
609597	C1865044	Disease or Syndrome	#N/A	#N/A	#N/A
300624	C0016667	Disease or Syndrome	#N/A	908	1-5 / 10 000
604290	C1858583	#N/A	Hemic and Lymphatic Diseases [mesh:C15]	48818	1-9 / 1 000 000
	Concept Unique Identifiers (CUI)			ORPHAnet ID	

OMIM DisGeNET orphanet

MeSH superclass annotation of rare genetic diseases

- > Top 10 MeSH terms	Count
congenital, hereditary, and neonatal diseases and abnormalities	1705
nervous system disease	945
musculoskeletal diseases	595
nutritional and metabolic diseases	549
pathological conditions, signs and symptoms	415
eye diseases	328
skin and connective tissue diseases	296
cardiovascular diseases	203
hemic and lymphatic diseases	182
female genital diseases and pregnancy complications	174

Which disease classes have been described when?

Conclusion: The creation of this dataset is an example how linking data can give benefit and allows drawing new conclusions – e.g. about the documentation of rare diseases and their causative genes. A crucial part is identifier and entity mapping, which allows to link data across different resources.

Funding statement: This work was funded by ELIXIR, the research infrastructure for life-science data (implementation study MolData2). FE and NBS are funded by Stichting Terre, the Dutch Rett syndrome funds. FE and CE are funded by EJP-RD.

Correspondence to:
friederike.ehrhart@maastrichtuniversity.nl
ORCID: 0000-0002-7770-620X
http://www.bigcat.unimaas.nl

Preprint on BioRXIV
https://www.biorxiv.org/content/biorxiv/early/2019/04/01/595819.full.pdf
Data on figshare
https://figshare.com/collections/Gene-Rare_Disease-Provenance_dataset_collection/4400798

Department of Bioinformatics-BIGCAT
Maastricht University, The Netherlands