Title

Open data informatics and data repurposing for IBD

Permalink
https://escholarship.org/uc/item/6rs8n1vj

Journal
Nature Reviews Gastroenterology & Hepatology, 15(12)

ISSN
1759-5045

Authors
Rudrapatna, Vivek A
Butte, Atul J

Publication Date
2018-12-01

DOI
10.1038/s41575-018-0050-5

Peer reviewed

Published in final edited form as:
Nat Rev Gastroenterol Hepatol. 2018 December ; 15(12): 715–716. doi:10.1038/s41575-018-0050-5.

Open data informatics and data repurposing for IBD

Vivek A. Rudrapatna[1,2], **Atul J. Butte**[2,3,*]

[1]Division of Gastroenterology, Department of Medicine, University of California, San Francisco, CA, USA.

[2]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA.

[3]Department of Pediatrics, University of California, San Francisco, CA, USA.

Abstract

Biomedical ‘big data’ has opened opportunities for data repurposing to reveal new insights into complex diseases. Public data on IBD have been repurposed for novel diagnostics and therapeutics, and these datasets continue to grow. Here, we discuss the practicalities and implications of open data informatics for IBD.

We live in a world of biomedical ‘big data’, in which the volume of data is so large and complex that our traditional methods of storage and analysis are challenged. Historically, the most common illustration of this concept has been by reference to nucleotide sequence data. For example, GenBank, the open-access genetic sequence repository of the US National Center for Biotechnology Information (NCBI), has been doubling in size approximately every 18 months since its inception in 1982. The drivers underlying the exponential growth of sequencing data are multiple — continued innovation in computing and digital storage, the declining costs of sequencing and the extension of technology into the microbial and mammalian single-cell levels.

The past 5–10 years have seen biomedical data expand into non-sequence related domains. In the USA, governmental legislation has been an essential force driving the digitization of medical records. By necessity, the subsequent explosion of electronic health record data simultaneously led to the birth of a new field at the intersection of information technology and health-care known as ‘clinical informatics’. In a short time this field has expanded into other platforms, including the patient blogosphere, as well as personalized devices that track sleeping patterns, bowel habits and everything in between. Although much of this data has historically been siloed in the private domain, increasing proportions of it are finally becoming public. This development has been important for many reasons, not least of which is the problem of scientific reproducibility. Indeed, the past several decades have seen an

[*] atul.butte@ucsf.edu.

Competing interests
The authors declare no competing interests.

Supplementary information
Supplementary information is available for this paper at https://doi.org/10.1038/s41575-018-0050-5.

explosion of systematic reviews and meta-analyses adjudicating on conflicting results and proposing field-wide consensuses.

However, we and others are most excited about public data for an entirely different reason: data repurposing. That is, intersecting and slicing datasets along axes orthogonal to those of the original study to unlock novel insights. For example, prior work from our group and others computationally identified a marker for diabetes-associated pancreatic cancer by intersecting published pancreatic cancer gene expression datasets with databases of validated type 2 diabetes mellitus susceptibility genes; the resultant hit (*FABP1*) was subsequently validated by tissue immunohistochemistry[1]. We used a similar approach of joining public gene expression studies with serum proteomic analysis using protein electrophoresis to identify novel biomarkers for preeclampsia, which were subsequently validated by ELISA[2]. Data repurposing studies performed by our group have also used clinical data, including that from The National Health and Nutrition Examination Survey (NHANES), a programme of the US Centers for Disease Control. Although this annual survey was originally designed to systematically characterize the health and nutritional status of Americans, it has been combined with complementary datasets such as the International Study of Macro/ Micronutrients and Blood Pressure (INTERMAP) to identify micronutrient modulators of blood pressure[3]. NHANES data have also been used as the basis of the first environment-wide association study (EWAS) to identify new chemical mediators of diabetic risk[4].

These examples imply that data repurposing can be most effective when applied across multiple dense and complementary datasets, as might be available for any well-studied complex disease. For this reason, we see this approach as being promising for the future of IBD research. IBD is a complex disease entity with a pathological footprint seen at virtually every level of characterization: genome, transcriptome, microbiome, metabolome, immunome and exposome. Our survey of the public domain confirms that the size and richness of the available data on IBD is tremendous (Table 1; full version in Supplementary Table 1), and this public data will continue to expand.

Despite the complexity and heterogeneity of the disease, we know that IBD is susceptible to the methods of data repurposing because we and others have seen it work before. For example, one study performed a re-analysis of NHANES data and identified the compound 4-tert-octylphenol as a possible environmental toxin associated with ulcerative colitis[5]. Work published by a different group applied cellular deconvolution techniques to previously published gene expression datasets to identify cell-level and gene-level signatures that distinguish anti-TNF therapy non-responders from responders[6]. This analysis identified differences in plasma cell and inflammatory macrophage counts, which were subsequently validated by immunohistochemistry of tissue biopsy samples from patients with IBD. Their approach also identified a single gene expression signature in peripheral blood (*TREM1*) that identifies primary non-responders with an area under the receiver-operator curve of 94%[6]. Lastly, our group used public gene expression data to compare disease signatures for IBD against those measured by drug screening. This work identified the anti-epileptic agent topiramate, which was subsequently shown to reverse an animal model of chemically induced colitis[7].

Data repurposing has shown early promise for the study of IBD aetiology, prognosis, and treatment, but we believe that the best is yet to come. For example, not only does the vast majority of structured data go unanalysed, but public data continues to grow rapidly and at a rate outpacing the influx of data scientists. Our survey, summarized in Supplementary Table 1. reveals new categories of public IBD data that did not exist even 5 years ago, such as the immunome[8,9] and exposome. Wearables, apps, and the 'Internet of Things' promise to expand these categories of available data further. There is enormous room for newcomers to enter the field and to have an impact.

In addition, our informatics toolkit has never been richer. Breakthroughs in machine learning are starting to enable personalized medicine in ways never before imagined. Natural language processing is also becoming increasingly sophisticated and promises to help unlock the 'dark matter' of unstructured data such as notes and reports. One notable barrier to these approaches in general is privacy concerns. Nevertheless, some groups have already succeeded in de-identifying their notes for public consumption[10], and more are sure to follow. We are optimistic that these tools will continue to grow in potential, releasing increasing quantities of valuable clinical content into the open domain.

Finally, the most compelling rationale for our advocacy of data repurposing is that it can be done by anyone with little more than a computer and Internet access. Of course, all computationally-identified targets require experimental validation to be trustworthy. But short of this process, and especially for an increasingly global disease, data repurposing is a method that can scientifically engage everyone, everywhere.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

V.A.R. was supported by the NIH training grant T32 DK007007–42. A.J.B. was supported by the National Institute of Allergy and Infectious Diseases (Bioinformatics Support Contract HHSN316201200036W).

References

1. Sharaf RN et al. Computational prediction and experimental validation associating FABP-1 and pancreatic adenocarcinoma with diabetes. BMC Gastroenterol. 11, 5 (2011). [PubMed: 21251264]
2. Liu LY et al. Integrating multiple 'omics' analyses identifies serological protein biomarkers for preeclampsia. BMC Med. 11, 236 (2013). [PubMed: 24195779]
3. Tzoulaki I et al. A nutrient-wide association study on blood pressure. Circulation 126, 2456–2464 (2012). [PubMed: 23093587]
4. Patel CJ, Bhattacharya J & Butte AJ An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLOS ONE 5, e10746 (2010). [PubMed: 20505766]
5. de Silva PS et al. Association of urinary phenolic compounds, inflammatory bowel disease and chronic diarrheal symptoms: Evidence from the National Health and Nutrition Examination Survey. Environ. Pollut 229, 621–626 (2017). [PubMed: 28689150]
6. Gaujoux R et al. Cell-centred meta-analysis reveals baseline predictors of anti-TNFα non-response in biopsy and blood of patients with IBD. Gut. 10.1136/gutjnl-2017-315494 (2018).
7. Dudley JT et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. Sci. Transl. Med 3, 96ra76 (2011).

8. Bhattacharya S et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci. Data 5, 180015 (2018). [PubMed: 29485622]

9. Zalocusky KA et al. The 10,000 Immunomes Project: a resource for human immunology. bioRxiv. 10.1101/180489 (2017).

10. Johnson AEW et al. MIMIC-III, a freely accessible critical care database. Sci. Data 3, 160035 (2016). [PubMed: 27219127]

Table 1 |

Key public repositories for IBD research

Resource name	IBD-related records	Notes	URL
Patient-generated clinical data			
IBD Partners	15,000 patients	Detailed patient surveys since 2011, easy access	ccfa.med.unc.edu
Genome and gene expression			
CWAS Catalog	60 studies, 1,522 associations	Most widely used, public and manually-curated GWAS database	ebi.ac.uk/gwas/
Sequence Read Archive (SRA)	6,700 sequences	Lists raw genomic or transcriptomic data of patients and their microbiome	ncbi.nlm.nih.gov/sra
Database of Genotypes and Phenotypes (dbGaP)	10 studies	Some overlap with SRA or GEO but has additional patient phenotype information	ncbi.nlm.nih.gov/gap
UK10K	4,000 control genomes	Source of control genomes for prospective IBD gene identification	uk10k.org
Gene Expression Omnibus (GEO)	~3,500 samples	NCBI gene expression database	ncbi.nlm.nih.gov/gds
Integrated or multiomics			
IBD-MDB	2,500 samples, 4,500 files	Multidimensional microbiome characterization of well-phenotyped multi-centre patient cohorts	ibdmdb.org
UK Biobank	~2,000 Crohn’s disease, ~3,800 ulcerative colitis samples	Genotype, questionnaire data, biometrics, some hospitalization and cancer or death outcomes	ukbiobank.ac.uk/

Full version available in Supplementary Table 1.