Supplementary Data

The Optum™ Impact™ National Managed Care Benchmark Database is a comprehensive, de-identified U.S. healthcare claims database that is representative of the non-elderly (65 years old and younger), insurance-carrying US population. The data are collected from 46 healthcare insurance providers serving members across nine census regions since 1997. By 2011, the database had accumulated claim records for 105.2 million insured members, based on approximately 15,000 ICD9 disease or procedure codes (abbreviated for the International Classification of Diseases, 9th Edition, published by the World Health Organization, WHO). We chose to study the claims from 2000 to 2011 for stabilized data volume and quality. We then adopted the ICD9 grouping methods developed by Denny et al [1] in their PheWAS study to group all ICD9 codes into approximately 1,700 meaningful clinical categories. This allowed us to calculate disease prevalence, treatment cost, and our newly defined Health Research Opportunity Index (Health ROI).

University of Chicago Hospital Data comprises de-identified medical records representing both outpatients and inpatients from the South Chicago area in 2011. It represents 146,989 unique patients; among them 120,303 (81.8%) are at the age of 65 or under and 26,686 (17.8%) are older than 65. The summary statistics of the sampled population is listed in Supplemental Table 3. The kernel density plot in Figure 2B shows that the ROIs computed for older and younger populations are very tightly correlated (Spearman's rank correlation coefficient equals 0.87 with p value ~0). Kernel density estimation is a non-parametric data smoothing technique where inferences about the population are made, based on a finite data sample [2, 3] and is implemented using Seaborn statistical data visualization package (http://stanford.edu/~mwaskom/software/seaborn/index.html).

The National (Nationwide) Inpatient Samples (NIS) is the largest all-payer inpatient care database in the United States. Each year of the NIS provides information on approximately 8 million inpatient stays from about 1,000 hospitals participating in the Healthcare Cost and Utilization Project (HCUP). To compare with the University of Chicago Hospital Data and the Optum Database, we analyzed the NIS data for 2011 too. Summary statistics for the NIS population are listed in Supplemental Table 4 and the kernel density plot is shown in Figure 2: The ROIs for patients of 65 and below and those over 65 correlate at *0.59* with *p* value less than 10^{-124}. NIS data does not contain complete information about annual treatment costs. So when calculating health ROI from NIS data, we used disease incidence in the NIS and average annual treatment costs from Optum, based on the strong but reasonable assumption that it costs the same price to treat disease in those under and over 65.

The World Health Organization (WHO) data: The WHO publishes global burden of diseases based on the number of deaths, Years Lost due to Disability (YLD), Years of Life Lost (YLLs), and Disability-Adjusted Life Year (DALYs) since 1990 [4]. We compared the latest burden of disease estimates in the United States published by WHO (2010) alongside our insurance and hospital-based burden of disease estimates (e.g., prevalence and treatment cost) in the same year from *The Optum™ Impact™ National*

Managed Care Benchmark Database for the 93 disease categories in both. Results are summarized in Table 1 and discussed in the main manuscript.

MEDLINE/PubMed data: We downloaded all Medical Subject Heading (MeSH)-indexed abstracts in English from the MEDLINE database (via the PubMed query interface), the biomedical publication database maintained by the United States National Library of Medicine (NLM) for each year from 2000 until 2011. Then we used the number of publications annotated by each Medical Subject Heading (MeSH) disease term to approximate attention received from the biomedical research community in each disease area. Our assumption was that total summary numbers most closely reflect the focus and attention of the entire research community. We also performed an additional experiment that restricted this sample to only research oriented articles, including original journal articles, clinical trials, and meta-analyses. We find that results from this more limited sample are broadly the same. As research articles take a dominant share in MEDLINE publications, the correlation between English *research* articles and all English publications related to specific diseases is higher than 0.999 for each year between 2000 and 2011 (see Supplemental Table 2). Therefore we kept other publication types, such as books, case reports, evaluation studies, guidelines, technical reports and lectures in the analysis. In some cases, those non-research publication types have a bigger impact on individual scientists and healthcare practitioners.

Clinical trial database: clinicaltrial.gov is a mandated clinical trial registration maintained by the United States National Library of Medicine. We downloaded its aggregated, MeSH-indexed database extract from the website of Clinical Trials Transformation Initiative (http://www.ctti-clinicaltrials.org). Similarly we used the number of clinical trials related to a specific disease to approximate feasibility and popularity of carrying out clinical research in each disease area in a given year during 2000 and 2011. We indexed these trials in our analysis by trial start date.

NIH funding information for 83 disease categories. The U.S. National Institutes of Health (NIH) publishes its annual support level since 2009 for various research, condition, and disease categories based on grants, contracts, and other funding mechanisms at http://report.nih.gov/categorical_spending.aspx. We identified the same data for the Fiscal Year 2003- 2008 in earlier NIH publications and manually mapped 83 clearly defined diseases according to the PheWAS groupings of ICD9 codes [1]. We did not include additional NIH support funded from the American Recovery and Reinvestment Act accounts in that calculation, as it was only one-time support for the years 2009 and 2010.

Mapping disease terminologies. The data sources described above use different terminologies to annotate disease. In order to facilitate the integration analysis, we further mapped the PheWAS and MeSH disease names via ICD9, according to a refined method developed by *J. Cimino* and colleagues [5]. Not all PheWAS terms were mapped to MeSH terms: 371 out of 1,722 (or 21.5%) were excluded. Major reasons include (1) some PheWAS terms are not diseases or phenotypes strictly speaking, such as ‘chemotherapy’, ‘radiotherapy’ and ‘liver replaced by transplant’; (2) MeSH does not cover diseases at the same granularity as ICD9 and PheWAS; terms such as ‘Diabetes

type 2 with ketoacidosis or uncontrolled diabetes' and 'lupus' are missing from MeSH, and (3) some terms were not mapped due to imperfections in the UMLS mapping and semantic types.

Supplementary Method: Theoretical Method for Estimating the Uncertainty/Variance in Health ROI

$$ROI_d = \log_{10}\left(\prod_{m\in\{M-b\}} \frac{X_{bd}}{X_{md}}\right).$$

Using the Delta method,

$$\begin{aligned} Var(ROI_d) \backsimeq\ & Var(X_{bd})\left[\frac{\partial ROI_d}{\partial X_{bd}}\right]^2 + \sum_m Var(X_{md})\left[\frac{\partial ROI_d}{\partial X_{md}}\right]^2 \\ & + 2\sum_m \frac{\partial ROI_d}{\partial X_{bd}}\frac{\partial ROI_d}{\partial X_{md}} Cov(X_{bd}, X_{md}) \\ & + 2\sum_{\substack{n\in(M-b),\\ m\in(M-b), n\neq m}} \frac{\partial ROI_d}{\partial X_{nd}}\frac{\partial ROI_d}{\partial X_{md}} Cov(X_{nd}, X_{md}), \end{aligned}$$

$$\frac{\partial ROI_d}{\partial X_{bd}} = \frac{|M|-1}{X_{bd}\ln(10)}.$$

In our implementation of the Health ROI, M includes three factors, namely disease burden, literature and clinical trials. Therefore,

$$\frac{\partial ROI_d}{\partial X_{bd}} = \frac{2}{X_{bd}\ln(10)}.$$

Similarly,

$$\frac{\partial ROI_d}{\partial X_{md}} = -\frac{1}{X_{md}\ln(10)}.$$

$$X_{bd} = p_{bd}C_{bd}.$$

where C_{bd} is an estimated treatment cost burden of condition *d* and p_{bd} is the relative prevalence or probability of condition *d.* Thus the variance of the binomially distributed random variable X_{bd} would be

$$Var(X_{bd}) \cong \frac{C_{bd}^2 p_{bd}(1-p_{bd})}{N}$$

where N is the total number of unique patients described in the dataset and the constant C_{bd} represents the mean treatment cost associated with a given condition.

Finally, $Var(X_{md})$ can be estimated using bootstrapping, and $Cov\left(X_{bd}, X_{md}\right) = \rho_{bm}\sqrt{Var(X_{bd})Var(X_{md})}$, where ρ_{bm} is an empirical correlation. Because of the very large sample sizes used in our analysis, the variances of Health ROIs are vanishingly small and thus are not shown in the paper.

Supplementary Table 1: Correlation between NIH funding in 83 disease areas and 4 metrics, namely disease burden (measured by total treatment cost in million population), publications, clinical trials, and health ROI.

	Disease Burden				Publications			
Year	Pearson correlation coefficient	*p*-value	Spearmen rank coefficient	*p*-value	Pearson correlation coefficient	*p*-value	Spearmen rank coefficient	*p*-value
2011	0.019	0.865	0.100	0.382	*0.296*	*0.008*	*0.677*	*0.000*
2010	0.029	0.800	0.069	0.545	*0.292*	*0.009*	*0.597*	*0.000*
2009	0.031	0.783	0.102	0.370	*0.294*	*0.009*	*0.606*	*0.000*
2008	0.013	0.909	0.087	0.446	*0.299*	*0.008*	*0.587*	*0.000*
2007	0.028	0.803	0.066	0.566	*0.338*	*0.002*	*0.600*	*0.000*
2006	0.036	0.754	0.051	0.655	*0.347*	*0.002*	*0.606*	*0.000*
2005	0.039	0.735	0.057	0.622	*0.350*	*0.002*	*0.636*	*0.000*
2004	0.045	0.692	0.037	0.745	*0.356*	*0.001*	*0.607*	*0.000*
2003	0.062	0.592	0.056	0.629	*0.360*	*0.001*	*0.653*	*0.000*

	Clinical Trials				Health ROI[1]			
Year	Pearson Correlation Coefficient	*p*-value	Spearmen Rank Coefficient	*p*-value	Pearson Correlation Coefficient	*p*-value	Spearmen Rank Coefficient	*p*-value
2011	*0.275*	*0.014*	*0.501*	*0.000*	*-0.223*	*0.048*	*-0.302*	*0.007*
2010	*0.294*	*0.009*	*0.482*	*0.000*	-0.198	0.080	*-0.253*	*0.024*
2009	*0.293*	*0.009*	*0.440*	*0.000*	-0.181	0.110	*-0.283*	*0.011*
2008	*0.310*	*0.005*	*0.439*	*0.000*	-0.213	0.060	*-0.245*	*0.030*
2007	*0.325*	*0.003*	*0.486*	*0.000*	*-0.225*	*0.047*	*-0.300*	*0.007*
2006	*0.320*	*0.004*	*0.480*	*0.000*	*-0.241*	*0.032*	*-0.344*	*0.002*
2005	*0.315*	*0.005*	*0.530*	*0.000*	*-0.238*	*0.036*	*-0.378*	*0.001*
2004	*0.331*	*0.003*	*0.461*	*0.000*	*-0.234*	*0.038*	*-0.347*	*0.002*
2003	*0.334*	*0.003*	*0.549*	*0.000*	*-0.253*	*0.025*	*-0.376*	*0.001*

[1]Health ROI was calculated using total treatment cost per million population, disease-specific publications and clinical trials.

Supplementary Table 2: Research articles dominate MEDLINE database

Year	All MEDLINE Articles in English	Research Articles in English	Disease specific correlation (Pearson's coefficient)
2000	474,286	456,446	0.9997255
2001	488,230	467,303	0.9996532
2002	504,244	483,230	0.9996817
2003	531,638	509,940	0.9996588
2004	574,511	550,694	0.9997459
2005	633,146	608,295	0.9997463
2006	678,224	654,512	0.9997636
2007	716,287	693,163	0.9997397
2008	765,184	740,746	0.9997406
2009	806,262	781,733	0.9996641
2010	869,674	845,651	0.9996908
2011	943,258	921,154	0.9996521

Supplementary Table 3: Summary Statistics of University of Chicago Hospital Data

	<=65		>65		All Ages	
Gender	**N**	**% (σ)**	**N**	**% (σ)**	**N**	**% (σ)**
Female	67,964	56.49 (0.14)	15,213	57.01 (0.30)	83,177	56.59 (0.13)
Male	52,332	43.50 (0.14)	11,473	42.99 (0.30)	63,805	43.41 (0.13)
Unknown	7	0.01 (0.00)	0	0.00 (0.00)	7	0.00 (0.00)
Age groups (years)	**N**	**% (σ)**	**N**	**% (σ)**	**N**	**% (σ)**
0-17	33,185	27.58 (0.13)	-	-	33,185	22.58 (0.11)
18-33	31,191	25.93 (0.13)	-	-	31,191	21.22 (0.11)
34-49	26,672	22.17 (0.12)	-	-	26,672	18.15 (0.10)
50-65	29,255	24.32 (0.12)	-	-	29,255	19.90 (0.10)
66-72	-	-	10,381	38.90 (0.30)	10,381	7.06 (0.07)
73-79	-	-	8,060	30.20 (0.28)	8,060	5.48 (0.06)
80-86	-	-	5,174	19.39 (0.24)	5,174	3.52 (0.05)
>86	-	-	3,071	11.51 (0.20)	3,071	2.09 (0.04)
Disease Name	**N**	**%**	**N**	**%**	**N**	**%**
Essential hypertension	17,544	14.58	17,195	64.44	34,739	23.63
Type 2 diabetes	6,431	5.35	6,954	26.06	13,385	9.11
Asthma	10,366	8.62	1,892	7.09	12,258	8.34
Depression	6,218	5.17	3,203	12.00	9,421	6.41
Benign neoplasm of colon	3,683	3.06	4,281	16.04	7,964	5.42
Congestive heart failure	3,008	2.50	4,093	15.34	7,101	4.83
Insomnia	1,856	1.54	1,200	4.50	3,056	2.08
Alzheimer's disease	42	0.04	884	3.31	926	0.63
Autism	330	0.27	1	0.00	331	0.23
Chlamydia	102	0.09	0	0.00	102	0.07

Supplementary Table 4: Summary Statistics of National Inpatient Samples Data

	<=65		>65		All Ages	
Gender	**N**	**% (σ)**	**N**	**% (σ)**	**N**	**% (σ)**
Female	3,069,299	56.91 (0.03)	1,572,457	58.48 (0.02)	4,641,756	57.94 (0.02)
Male	2,178,706	43.08 (0.03)	1,190,265	41.51 (0.02)	3,368,971	42.05 (0.02)
Unknown	436	0.00 (0.00)	129	0.01 (0.00)	565	0.01 (0.00)
Age groups (years)	**N**	**% (σ)**	**N**	**% (σ)**	**N**	**% (σ)**
0-17	1,182,494	22.53 (0.02)	-	-	1,182,494	14.76 (0.01)
18-33	1,233,415	23.50 (0.02)	-	-	1,233,415	15.40 (0.01)
34-49	1,126,315	21.46 (0.02)	-	-	1,126,315	14.06 (0.01)
50-65	1,706,217	32.51 (0.02)	-	-	1,706,217	21.30 (0.01)
66-72	-	-	813,070	29.43 (0.03)	813,070	10.15 (0.01)
73-79	-	-	741,313	26.83 (0.03)	741,313	9.25 (0.01)
80-86	-	-	702,480	25.43 (0.03)	702,480	8.77 (0.01)
>86	-	-	505,988	18.31 (0.02)	505,988	6.32 (0.01)
Disease Name	**N**	**%**	**N**	**%**	**N**	**%**
Essential hypertension	8,715	0.166	6,458	0.234	15,173	0.189
Type 2 diabetes	21,976	0.419	17,484	0.633	39,460	0.493
Asthma	3,096	0.059	976	0.035	4,072	0.051
Depression	20,311	0.387	1,187	0.043	21,498	0.268
Benign neoplasm of colon	3,578	0.068	3,709	0.134	7,287	0.091
Congestive heart failure	16,934	0.323	43,652	1.580	60,586	0.756
Insomnia	32	0.001	33	0.001	65	0.001
Alzheimer's disease	580	0.011	12,692	0.459	13,272	0.166
Autism	657	0.013	2	0.000	659	0.008
Chlamydia	48	0.001	0	0.000	48	0.001

Supplementary Table 5: Top 50 over/under studied conditions in 2011

Top 50 over studied conditions in 2011	
Name	**Short Name**
Breast cancer	breast cancer
Cervical cancer and dysplasia	cervical cancer dysplasia
Other symptoms of respiratory system	other Sx of respiratory system
Renal failure	renal failure
Cardiac arrest and ventricular fibrillation	cardiac arrest & VF
Ischemic Heart Disease	IHD
Cardiac dysrhythmias	arrhythmia
Peptic ulcer (excl. esophageal)	peptic ulcer
Male infertility and abnormal spermatozoa	male infertility & abnormal sperm
Cholelithiasis and cholecystitis	cholelithiasis & cholecystitis
Benign mammary dysplasias	mammary dysplasias
Otitis media and Eustachian tube disorders	otitis media & ETD
Gastrointestinal hemorrhage	GI bleeding
Gastritis and duodenitis	gastritis/duodenitis
Inflammatory diseases of prostate	inflammatory disease of prostate
Intestinal malabsorption	GI malabsorption
Cardiomyopathy	cardiomyopathy
Nephritis and nephropathy without mention of glomerulonephritis	membranoprolif nephr NOS
Chronic liver disease and cirrhosis	chronic liver disease/cirrhosis
Iron deficiency anemias	iron deficiency anemia
Cystitis and urethritis	cystitis & urethritis
Proteinuria	proteinuria
Peripheral vascular disease	PVD
Poisoning by antifungal antibiotics	Pois-antifungal antibiot
Disorders of carbohydrate transport and metabolism	carbohydrate transport & metabolism DO
Nephritis; nephrosis; renal sclerosis	nephritis/nephrosis /renal sclerosis
Gout and other crystal arthropathies	gout & other crystal arthropathies
Cancer of bone and connective tissue	sarcomas
Erythematous conditions	erythematous conditions
Osteomyelitis	osteomyelitis
Sepsis and SIRS	SIRS
Suicidal ideation or attempt	suicidal ideantion or attempt
Chronic ulcer of skin	chronic skin ulcer
Poisoning by hormones and synthetic substitutes	poisoning hormon NEC/NOS
Other disorders of pancreatic internal secretion	pancreatic DO NEC
Diabetes mellitus	diabetes
Psoriasis and related disorders	psoriasis & related DO
Disorders of protein plasma/amino-acid transport and metabolism	amino acid transport & metabolism DO

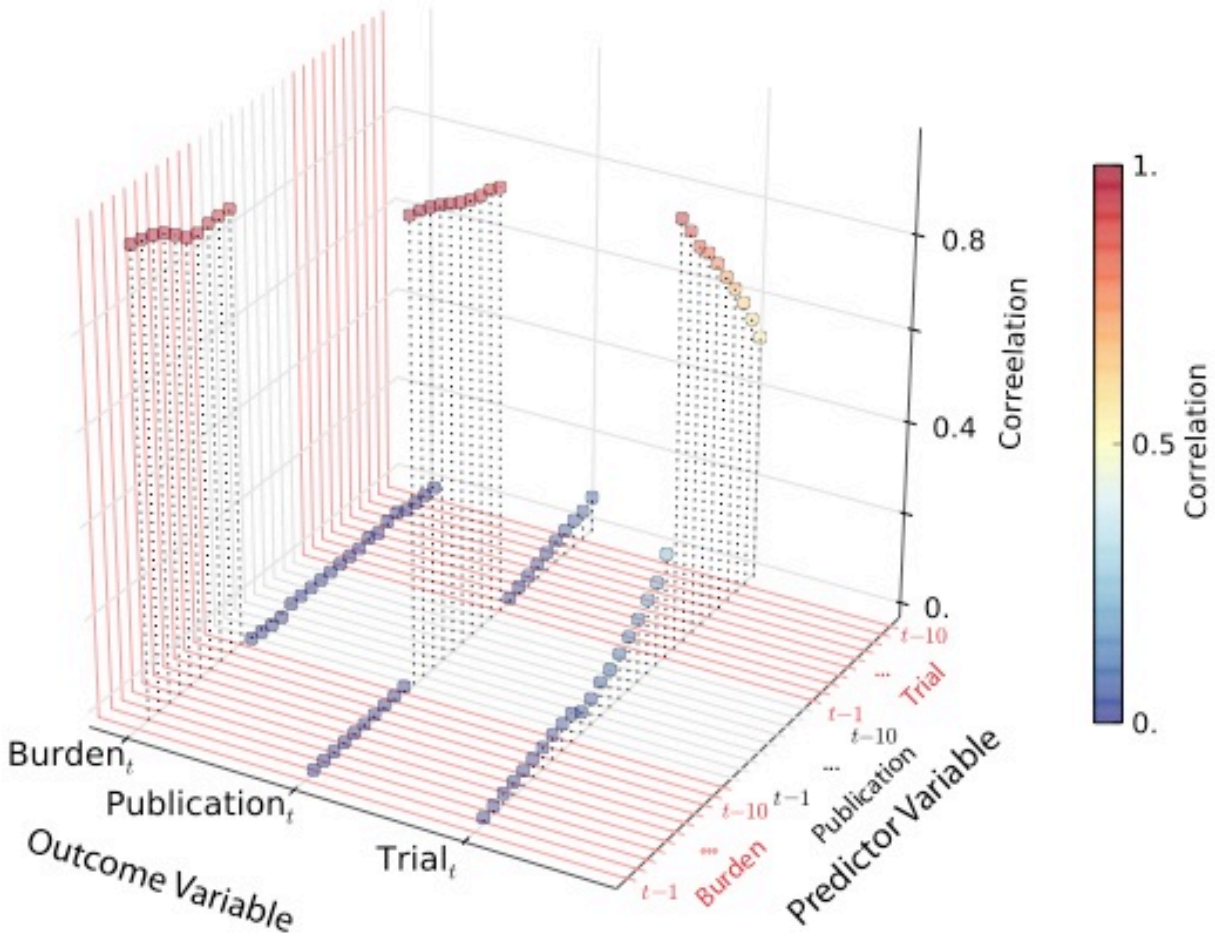
Adverse effects of antibacterials (not penicillins)	antibiotics side effect
Disorders of lipoid metabolism	lipid metabolism DO
Osteoarthrosis	OA
Nephritis and nephropathy in diseases classified elsewhere	Nephritis NOS in oth dis
Diverticulosis and diverticulitis	diverticulosis & diverticulitis
Sulfonamides	sulfonamides side effects
Atrial fibrillation and flutter	atrial fibrillation or flutter
Hypertension	hypertension
Injury to other and unspecified nerves	nerve injury NEC
Other biliary tract disease	other biliary tract disease
Lung disease due to external agents	lung disease due to external agents
Secondary diabetes mellitus	secondary diabetes

Top 50 under studied conditions in 2011

Name	Short Name
Chronic lymphocytic thyroiditis	hashimoto's thyroiditis
Other disorders of cervical region	other cervical DO
Palpitations	palpitation
Secondary malignancy of bone	secondary bone cancer
Mixed hyperlipidemia	hyperlipidaemia
Septal Deviations/Turbinate Hypertrophy	septal deviations/turbinate hypertrophy
Fluid overload	hypervolemia
Adjustment reaction	adjustment reaction
Cervical radiculitis	cervical radiculitis
Diarrhea	Diarrhea
Other signs and symptoms involving emotional state	Nervousness
Irregular menstrual cycle/bleeding	irregular period
Joint effusions	joint effusion
Galactorrhea	Galactorrhea
Contracture of joint	joint contracture
Seborrheic keratosis	seborrhoeic wart
Testicular hypofunction	testicular hypofunction
Dysuria	Dysuria
Symptoms involving female genital tract	Sx of female genital tract
Abnormality of gait	gait abnormality
Elevated sedimentation rate	elevated sed rate
Umbilical cord complications during labor and delivery	umbilical cord complications at labor
Hammer toe (acquired)	hammer toe
Costochondritis	Costochondritis
Muscle weakness	Myasthenia
Polydipsia	Polydipsia
Multiple gestation	multiple birth

Diverticulitis	Diverticulitis
Sebaceous cyst	sebaceous cyst
Acne	Acne
Changes in skin texture	skin texture change
Postmenopausal atrophic vaginitis	atrophic vaginitis
Colles' fracture	colles fracture
Otalgia	Otalgia
Ascites (non malignant)	Ascites
Voice disturbance	voice disturbance
Hyperglyceridemia	Hyperglyceridemia
Normal delivery	normal delivery
Lipoma of skin and subcutaneous tissue	lipoma of skin & subcutaneous tissue
Sacroiliitis NEC	Sacroiliitis
Other dyschromia	other dyschromia
Chronic inflammatory pelvic disease	chronic PID
Hematuria	Hematuria
Other persistent mental disorders due to conditions classified elsewhere	Mental disor NEC
Abnormal results of function study of liver	abn LFT
Other disorders of the kidney and ureters	other kidney/ureter DO
Calculus of ureter	ureteric calculus
Microscopic hematuria	microscopic hematuria
Heart transplant/surgery	Hrt dis postcardiac surg

Supplementary Figure 1: Dependence of burden of disease (measured by total treatment cost per million population) at year *t* (Burden$_t$), number of disease-specific publications published during year *t* (Publication$_t$), and disease-specific clinical trials initiated during year *t* (Trial$_t$) on values of these three quantities in the previous year (*t*-1), two years earlier (*t*-2), …, ten years earlier (*t*-10).



Supplementary Reference:

1. Denny, J.C. et al. *Bioinformatics* **26**, 1205-1210 (2010).
2. Parzen, E. *The annals of mathematical statistics*, 1065-1076 (1962).
3. Rosenblatt, M. *The Annals of Mathematical Statistics* **27**, 832-837 (1956).
4. Murray, C.J. & Lopez, A.D. *N Engl J Med* **369**, 448-457 (2013).
5. Cimino, J.J., Johnson, S.B., Peng, P. & Aguirre, A. *Proc Annu Symp Comput Appl Med Care*, 730-734 (1993).