RESEARCH PROTOCOL:
Time trends in prevalence, incidence and survival of cancer in the OHDSI Network

Version: 3.0

Date: 19th November 2024

Acknowledgement: The analysis is based in part on work from the Observational Health Sciences and Informatics collaborative. OHDSI (http://ohdsi.org) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.

Table of contents

1 List of Abbreviations

AAPC	Average Annual Percent Change
APC	Annual Percent Change
CDM	Common Data Model
COPD	Chronic Obstructive Pulmonary Disease
EHDEN	European Health Data and Evidence Network
HIV	Human Immunodeficiency Virus
HPV	Human Papillomavirus Infection
IRB	Institutional Review Board
IR	Incidence Rate
IRR	Incidence Rate Ratio
OMOP	Observational Medical Outcomes Partnership
OHDSI	Observational Health Data Science and Informatics
PR	Prevalence Rate
SIDIAP	Information System for Reseach in Primary Care
SNOMED	Systematized Nomenclature of Medicine
WHO	World Health Organization

2 Responsible Parties

2.1 Investigators

Investigator/Author	**Institution/Affiliation**
Talita Duarte-Salles*	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Edward Burn	University of Oxford
Asieh Golozar	Odysseus Data Services, Inc, Cambridge, MA USA, OHDSI Center at the Northeastern University, Boston, MA USA
Irene López Sánchez	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
Laura Pérez Crespo	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
Agustina Giuliodori Picco	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
Berta Raventós	Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
Anna Palomar Cros	Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
Anton Barchuk	Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
OHDSI Oncology WG	

*Principal Investigator

Authorship in scientific manuscripts will follow ICMJE authorship criteria (http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html). The Responsible Parties involved in this protocol take accountability for the overarching protocol, package development, providing assistance to sites running the analysis and ensuring site-specific governance is adhered to in all publications generated from this protocol.

3 Abstract

Objectives: The main aim of this project is to estimate time trends in prevalence and incidence rates, and short- and long-term survival of site-specific cancers in the OHDSI network.

Design: This study will be a multinational observational cohort study and will be conducted using a network of large real world data sources that have been mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

Setting: Population-based, electronic health records, claims and registry data from primary and secondary care.

Participants: Individuals with no prior history of cancer (for incidence and survival analyses only), and who have been on the database for at least 1 year before study entry.

Outcomes: Prevalent and incident cancer diagnoses and overall as well as 1-, 5-, and 10-year survival of site-specific cancers.

Data analyses: The OHDSI Cohort Diagnostics package will be used to assess the fitness of use of cancer data on each database. We will calculate prevalence (PR) and incidence rates (IR) with 95% confidence intervals (95%CI) for each year and study period by dividing the number of ever and first recorded cases of cancer, respectively, by 1,000 person-years of follow-up, overall and stratified by demographics and relevant comorbidities. The overall and 1-, 5-, and -10-year survival rates will be calculated as the percentage of people who have been diagnosed with cancer and are still alive during the study period as well as one or five years after diagnosis, respectively, per year and stratified by pre-defined subgroups. To assess the incidence trend over time, we will calculate the IRs in 5 year periods and then calculate the incidence rate ratios (IRRs) and their corresponding 95%CI to analyze the differences in incidence between the defined time periods.

4 Amendments and Updates

Number	Date	Section of study protocol	Amendment or update	Reason
Version 2.0	13/06/2023	All	Update	Update to incorporate team comments and suggestions
Version 3.0	19/11/2024	All	Amendment	Amendment to incorporate a new objective to the study

5 Milestones

Milestone	Planned date
Final version of study protocol	20/11/2024
Create phenotype definitions	30/11/2024
Study package release	30/01/2025

Data analyses	01/03/2025
Writing of scientific paper(s)	01/06/2025

6 Rationale and Background

Cancer is currently a leading cause of morbidity and mortality worldwide. Estimates from the World Health Organization (WHO) in 2019 showed that cancer is the first or second leading cause of death before the age of 70 years in 112 of 183 countries and ranks third or fourth in a further 23 countries.[1] The burden of cancer incidence and mortality is rapidly growing worldwide, reflecting both aging and growth of the population as well as changes in the prevalence and distribution of the main risk factors for cancer.[2,3] Worldwide, female breast cancer is the most commonly diagnosed cancer, followed by lung, colorectal, prostate, and stomach cancers, while lung cancer is the leading cause of cancer death, followed by colorectal, liver, stomach, and female breast cancers.[4] The continuous surveillance and monitoring of trends in cancer incidence and survival are needed for the development, implementation and evaluation of health policies aiming to reduce the burden of disease.

Multiple factors are attributed to cancer incidence and cancer survival. Some of the main cancer risk factors include age, race, smoking, alcohol consumption, cardiometabolic conditions (e.g.: cardiovascular diseases, obesity, hypertension, type 2 diabetes mellitus), infectious diseases (e.g.: hepatitis B/C, human immunodeficiency virus (HIV)), among others. [5,6] Survival rates are known to vary by cancer type as well as being influenced by cancer stage at detection and treatment. It can also depend on an individual's health, presence of comorbidities and other tumour-related factors. Recent advances in screening programmes and treatments have improved the survival rates for some cancer sites. [7] However, important disparities still exist in healthcare systems and countries. [7,8] Understanding the complexities of individual factors related to cancer incidence and survival, whilst investigating the population subgroups that are more prone to be at risk, is important for being able to plan future population-based interventions. [9]

Real world data from healthcare services promise to vastly expand clinical research, providing data for large-scale studies that would not be feasible with traditional research data collection methods. The Observational Health Data Sciences and Informatics (OHDSI; www.ohdsi.org) is a multi-stakeholder and interdisciplinary international network that generates open science through large-scale analytics of real world data. [10] This data network offers a unique opportunity to study cancer incidence and survival as it includes more than 100 healthcare databases from 20 countries including over 1.5 billion individual records with longitudinal data. OHDSI analyses are based around the use of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which is used to standardize otherwise disparate datasets. In this project, we will estimate time trends in prevalence, incidence rates and survival of cancers in the OHDSI network.

7 Study Objectives

The **main aim** of this project is to estimate time trends in prevalence, incidence rates and short- and long-term survival of site-specific cancers in the OHDSI network.

The **specific study objectives** include:
1) To estimate prevalence rates of site-specific cancers by calendar year, age, sex, and comorbidities.
2) To estimate incidence rates of site-specific cancers by calendar year, age, sex, and comorbidities.
3) To describe demographic, clinical characteristics and medications of individuals with cancer at the time of diagnosis.
4) To estimate overall, short- and long-term survival of site-specific cancers by calendar year, age, sex, and comorbidities.

8 Research Methods

8.1 Study Design and Study Period

This study will be a multinational observational cohort study. The study period will start on 1st January 2000 or one year after the earliest date of available data in each database and span up to the 31st December 2024 (or the last date of available data in each database).

8.2 Data Sources

The study will be conducted using a network of large real world data sources that have been mapped to the OMOP Common Data Model in collaboration with the OHDSI and European Health Data and Evidence Network (EHDEN) initiatives. The OMOP Common Data Model (https://github.com/OHDSI/CommonDataModel/wiki) includes a standard representation of health care experiences (such as information related to drug utilization and condition occurrence), as well as common vocabularies for coding clinical concepts, and enables consistent application of analyses across multiple disparate data sources.[10] The present study will be conducted in multiple databases in the OHDSI network willing to participate. Databases that have already committed to participate include: the Information System for Research in Primary Care (SIDIAP; Spain), the Integrated Primary Care Information (IPCI, The Netherlands), The Netherlands Cancer Registry (IKNL, The Netherlands), the Clinical Practice Research Datalink (CPRD, UK).

8.3 Study Population

All individuals registered in each of the data sources will be eligible for inclusion in the study. However, study participants will be required to have a year of prior history observed in the database before contributing observation time. Participants with prior history of a diagnosis of cancer (any, excluding non-melanoma skin cancer) any time prior to their index date will be excluded (for incidence and survival estimation).

For incidence rates, study participants will begin contributing person time on the respective date of the latest of the following: 1) study start date (1st January 2000 or the earliest date of available data in in each of the data sources), 2) date at which they have sufficient prior

history (defined as 365 days), 3) date at which they reach a minimum age (where age strata are being considered).
Participants will stop contributing person time at the earliest date of the following: 1) study end date (end of available data in each of the data sources), 2) date at which their observation period ends, 3) the last day in which they have the maximum age (where age strata are being considered). Where there are multiple age strata, study participants will contribute to each strata while they satisfy the conditions of that strata (i.e. when they reach the limit of one age strata they will begin contributing to the next).

8.4 Cancer prevalence, incidence and survival

Cancer prevalence and incidence will be defined as the occurrence for ever or the first time in the person's history, respectively. Conditions in the OMOP CDM use the Systematized Nomenclature of Medicine (SNOMED) as the standard vocabulary for diagnosis codes. The complete list of codes used to define each cancer site are available in Appendix 1.

For those databases with available data on date and/or cause of death, overall, 1-, 5-, and 10-year survival will be estimated. We will also report overall mortality, cancer-specific mortality, and other-cause mortality when possible.

8.5 Stratifications

Each target cohort will be analysed in full and stratified on sex (male, female), age (0-9; 10-19; 20-29; 30-39; 40-49; 50-59; 60-69; 70-79; 80-89; 90-99; 100 and over), smoking status, asthma, type 2 diabetes mellitus, hypertension, cardiovascular diseases, chronic obstructive pulmonary disease (COPD), obesity, human immunodeficiency virus (HIV), human papillomavirus infection (HPV), viral hepatitis, chronic kidney disease, autoimmune conditions, and depression. These stratifications will only be performed in those databases with available information on the variables listed above. The definition of these conditions can be found in Appendix 2. Female breast cancer will be stratified by menopausal status which will be defined based on age as pre- (less than 50 years) and post-menopausal (50 years or above). All stratum are pending meeting minimum reportable cell counts (as specified by data owners).

8.6 Other variables

Large-scale patient-level characterization will be conducted at the time of diagnosis. Medical history and medications use history will be assessed at any time prior cancer diagnosis and up to 366 days before index date, for 365 days to 31 days before index date, for 30 to 1 day before index date and at index date. We will also report medication use for 1 to 90 days post index-date.

A list of pre-specified comorbidities and medications will be described. These will include:

- Medical History: Anemia, Anxiety, Asthma, Arterial fibrillation, Cerebrovascular disease, Chronic kidney disease, Chronic liver disease, Chronic obstructive pulmonary disease (COPD), Coronary arteriosclerosis, Chron's Disease, Dementia, Depressive disorder, Gastro-oesophageal reflux disease (GERD), GI-Bleeding, Heart failure, Human

Immunodeficiency Virus (HIV), Human papillomavirus infection (HPV), Hyperlipidaemia, Hypertension, Hypothyroidism, Ischemic Heart Disease, Myocardial Infraction, Non-alcoholic Fatty-acid disease (NAFLD), Obesity, Osteoarthritis, Osteoporosis, Pancreatitis, Peripheral vascular disease, Pneumonia, Psoriasis, Pulmonary embolism, Renal impairment, Stroke, Type 1 Diabetes, Type 2 Diabetes, Ulcerative colitis, UTI, Venous thromboembolism, Viral Hepatitis

- Medications: Agents acting on the renin-angiotensin system, Antibacterials for systemic use, Antidepressants, Antiemetics, Antiepileptics, Anti-inflammatory and anti-rheumatic drugs, Antineoplastics, Anti-psoriatic, Antithrombotic agents, Antivirals HCV, antivirals HIV, Beta-blocking agents, Calcium channel blockers, Contraceptives, Diuretics, Drugs for acid related disorders, Drugs for obstructive airway diseases, Drugs used in diabetes, Immunosuppressants, Iron preparations, Lipid-modifying agents, Opioids, Psycholeptics and Psychostimulants.

8.7 Data quality assessment

Overall data quality of the OHDSI Network databases are assessed by each data partner using the data quality dashboard after data mapping. [10] We will assess the created phenotypes including cancer events using a comprehensive cohort characterization tool, the OHDSI's CohortDiagnostic package (https://ohdsi.github.io/CohortDiagnostics/). For any cohort and data source mapped to OMOP CDM, this package systematically generates incidence rates (stratified by age, gender, calendar year, and database), cohort characteristics (all comorbidities, drug use, procedures, health utilization) and the actual codes found in the data triggering the various rules in the cohort definitions. The CohortDiagnostics package works in two steps: 1) Generate the utilization results and diagnostics against a data source and 2) Explore the generated utilization and diagnostics in a user-friendly graphical interface R-Shiny app. These diagnostics provide a consistent methodology to evaluate cohort definitions/phenotype algorithms across a variety of observational databases, allowing to compare the overlap between alternative definitions. This can allow researchers and stakeholders to understand the heterogeneity of source coding for exposures and health outcomes as well as the impact of various inclusion criteria on overall cohort counts. Data quality of cancer diagnoses will be assessed by comparing the calculated incidence rates of events in a specific database with the incidence reported in national/regional cancer registries.

8.8 Analysis

All analyses will be performed using code developed for the OHDSI Methods library and executed by data owners using R and stratified by database. Each data partner will execute the study code against their database containing patient-level data and will then return the results set which will only contain aggregated data. The results from each of the contributing data sites will then be combined in tables and figures. The code for this study will be made publicly available at https://github.com/ohdsi-studies/.

To answer the study objectives, a diagnostic package, built off the OHDSI Cohort Diagnostics (https://ohdsi.github.io/CohortDiagnostics/) library, will be included in the base package as a preliminary step to assess the fitness of use of phenotypes on each database.

If a database passes cohort diagnostics (as described in section 8.6), the full study package will be executed.

To address **objectives 1 and 2**, we will first summarize the socio-demographics characteristics and overall follow-up of individuals included in the study, with counts and percentages for categorical variables and median and interquartile ranges (IQR) for continuous variables. Incidence rates (IR) with 95% CI will be calculated using the R package IncidencePrevalence (https://cran.r-project.org/web/packages/IncidencePrevalence/index.html). IR will be computed for each year and study period by dividing first-recorded diagnoses by person-years at risk for the whole population or pre-defined subgroups (section 8.5). The first recorded for each outcome category will be included as an incident episode. Each outcome will be assessed separately and therefore, individuals will be able to contribute to one the incidence counts for more than one outcome. To exclude prevalent cases from incidence calculations, individuals diagnosed with any cancer will not be considered as eligible incident cases in future cohorts after the date of their first diagnosis. To assess the incidence trend over time, we will calculate the IRs in 5 year periods and then calculated the incidence rate ratios (IRRs) and their corresponding 95% confidence intervals (95% CIs) to analyze the differences in incidence between the defined time periods (to be defined depending on data availability).

We will conduct a large-scale patient-level characterization to address **objective 3.** For this part of the analysis, we will use the package CohortCharacteristics (https://cran.r-project.org/web/packages/CohortCharacteristics/index.html). Essentially, we will use the function summariseLargeScaleCharacteristics, which is used to summarize the large-scale characteristics of a cohort table. Medical history and medications will be assessed for any time prior cancer diagnosis and up to 366 days before index date, for 365 days to 31 days before index date, for 30 to 1 day before index date and at index date (see appendix 2 and 3). Medications will also be assessed 1 to 90 days post-index date. We will also compare these characteristics in our cohorts of interest to the proportion of these diseases in a random subsample of SIDIAP.

To address **objective 4**, the 1-, 5-, and 10-year survival rates will be calculated as the percentage of people who have been diagnosed with cancer and are still alive one, five, or ten years after diagnosis, respectively, per year, overall and also stratified by pre-defined subgroups.
Survival will be calculated using data on time at risk of death from any cause and the Kaplan-Meier method. Results will be reported as plots of the estimated survival curves as well as the estimated probability of survival at years 1, 5, and 10. This analysis will be conducted only for databases that collect data on mortality. We will also perform Kaplan-Meier analyses to assess time from cancer diagnosis to cancer-specific survival, and other cause-mortality-free survival. KM curves will be further stratified by study periods (to be defined depending on data availability).

We will use joinpoint modeling to examine the overall trends in cancer incidence and survival (*to see if this can be implemented using the OMOP CDM*).[11] The model involves fitting a series of joined straight lines on a logarithmic scale to the trends in the annual rates. The

direction and magnitude of the resulting trends are described by the annual percent change (APC), the linear slope, across each line segment between two joinpoints. The average annual percent change (AAPC) summarizes the overall trend over time using a weighted average of the APCs within the specified period. The default maximum number of joinpoints allowed will be set at 4. In describing the change, the term increase or the term decrease will be used when the APC or the AAPC is statistically significant ($p < 0.05$); otherwise, the term stable was used.[12]

9 Sample Size and Study Power

Since this study will be undertaken using population-based data, we will include all patients meeting the eligibility criteria described above. No prior sample calculation will be performed but we will do feasibility counts before running the stratified analyses.

10 Strengths and Limitations

10.1 Strengths

The main strengths of this study are the sample size and the real-world nature of the data. Our study will provide data from broadly representative and heterogeneous settings and geographical regions. Another strength of this study is its longitudinal design including prospectively collected data spanning up to 2021. In addition, the use of pre-existing standardised analytics and tools across the OHDSI network will facilitate the development of the study across different data sources.

10.2 Limitations

One potential limitation of the present study is misclassification of cancer diagnosis and/or date of death. However, cancer diagnosis has been already validated in some of the participating databases, and additionally, we will use the OHDSI Cohort Diagnostics tool to assess quality of cancer incidence in the participating databases and only include high quality data. The lack and/or missingness of information on relevant stratification factors; such as race, smoking habits, alcohol consumption, stage of tumor at diagnosis or cancer treatment; might also be a limitation of the study. Although we already anticipate that stratification for these factors will not be possible in all participating databases, we will be able to do it in a few databases that are part of the OHDSI network.

11 Protection of Human Subjects

The study uses only de-identified data. Confidentiality of patient records will be maintained at all times. All data partners executing the study within their data sources will have received institutional review board (IRB) approval or waiver for participation in accordance with their institutional governance prior to execution. Data custodians will remain in full control of executing the analysis and packaging results. The study will be executed across a federated and distributed data network, where analysis code is sent to participating data partners and only aggregate summary statistics are returned, with no sharing of patient-level data between organizations. There will be no transmission of patient-level data at any time during

these analyses. Study packages will contain minimum cell count parameters to obscure any cells which fall below allowable reportable limits.

13 Plans for Disseminating and Communicating Study Results

Open science aims to make scientific research, including its data process and software, and its dissemination, through publication and presentation, accessible to all levels of an inquiring society, amateur or professional and is a governing principle of the present study. Open science delivers reproducible, transparent and reliable evidence. All aspects of study (except private patient data) will be open and we will actively encourage other interested researchers, clinicians and patients to participate.

We will share the study protocol with the OHDSI community for feedback. This protocol will link to open source code for all steps to generating the study results, which will also be made publicly available at data.ohdsi.org.

We will deliver presentations at scientific venues including the annual OHDSI symposium. We will also prepare scientific publications for international scientific peer-review journals. We will publish the results of this study following the International Committee of Medical Journal Editors (ICMJE) authorship guidelines, and will report the results following the appropriate Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist.

The main findings of this project will be shared with the general public through social media channels. With dedicated support from the OHDSI communications specialist, we will deliver regular press releases at key project stages, distributed via the extensive media networks of the study partners.

References

1. International Agency for Research on Cancer, GLOBOCAN 2018 accessed via Global Cancer Observator. Accessed March 21, 2023.
2. World Health Organization (WHO). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. WHO; 2020. Accessed January 10, 2022. who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death.
3. Gersten, O. & Barbieri, M. Evaluation of the Cancer Transition Theory in the US, Select European Nations, and Japan by Investigating Mortality of Infectious- and Noninfectious-Related Cancers, 1950-2018. *JAMA Netw. Open* **4**, e215322 (2021).
4. Gersten, O. & Wilmoth, J. R. The Cancer Transition in Japan since 1951. *Demogr. Res.* **7**, 271–306 (2002).
5. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
6. Stein, C. J. & Colditz, G. A. Modifiable risk factors for cancer. *Br. J. Cancer* **90**, 299–303 (2004).

7. Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–47 (2016).
8. Arnold, M. *et al.* Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. *Lancet Oncol.* **20**, 1493–1505 (2019).
9. Allemani, C. *et al.* Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet* **391**, 1023–1075 (2018).
10. Srinivasan, S. *et al.* Small is essential: importance of subpopulation research in cancer control. *Am. J. Public Health* **105 Suppl 3**, S371-373 (2015).
11. Observational Health Data Sciences and Informatics. The Book of OHDSI [Internet]. 2019. Accessed January 10, 2022. https://ohdsi.github.io/TheBookOfOhdsi/.
12. Kim, H. J., Fay, M. P., Feuer, E. J. & Midthune, D. N. Permutation tests for joinpoint regression with applications to cancer rates. *Stat. Med.* **19**, 335–351 (2000).
13. Culp, M. B., Soerjomataram, I., Efstathiou, J. A., Bray, F. & Jemal, A. Recent Global Patterns in Prostate Cancer Incidence and Mortality Rates. *Eur. Urol.* **77**, 38–52 (2020).

Appendix 1. Preliminary code list of cancer

Attached as a standalone document

Appendix 2. Preliminary code list of conditions

Conditions	Concept id (Included)	Concept id (Excluded)
Anemia	439777	434701, 43022052
Anxiety	441542	
Asthma	317009, 4235703, 4279553	
Autoimmune condition	4137275, 81893, 40484648, 201254, 4063582, 134442, 257628, 254443, 4035611, 4083556, 438688, 45772123, 4262578, 2107559, 2107558, 2107560, 80809, 4145240, 46273478, 81931, 140168, 4102493, 46270482, 432295, 2108721, 4105026, 4107913, 432893, 76685, 374919, 4105005, 255304, 441928, 4331739, 46272236, 46274082, 46269999, 135215, 4058299, 4232076, 46269952, 46273369, 435216, 2107561, 2107572, 46273442, 4334806, 4297650, 201606, 46273477, 46269889, 194992, 4116142, 443394, 75614	
Atrial fibrillation	313217	
Cardiovascular disease	4329847, 316139, 432923, 4148906, 42535425, 43530727, 43530674, 439847, 376713, 441874, 375557, 372924, 4132309, 4317150, 442289, 321042, 4048809	314666, 315295, 443454, 437894
Cerebrovascular disease	381591	
Chronic kidney disease	194385, 46271022, 192279, 4263367, 261071, 201313, 4103224, 193253, 195314, 192359, 45768812	45769152, 195289, 195737, 43530912, 37116834, 195014, 197930, 197320, 4066005
Chronic liver disease	4212540	
COPD	255573, 258780	
Coronary arteriosclerosis	317576	
Chron’s disease	201606, 46269889, 46269999	
Dementia	37312036, 37312035, 4041685, 37312031, 37312030, 35608576, 4092747, 4182210, 37311665, 4043378, 45765480, 45765477, 37311890, 37312577, 4059191	37116464, 37017549, 4244346, 377788, 372610, 37017247
Depressive disorder	440383	438727, 436665, 40481798, 435520, 4224940

GERD	318800	
GI-Bleeding	192671	
Heart Failure	316139	315295
HIV	4276586, 44783356, 439727	4013105, 432554
HPV	40480043, 4084948, 4116193, 4080771, 4294441, 4291601, 619210, 760929, 4219870, 4266804, 4269876, 4084816, 44810559, 4304732, 4164483, 4129543, 4175989, 198075, 4080330, 4147672, 4078931, 4291605, 37116426, 4347555, 4177636, 4345817, 4148102, 4345473, 36713662, 37206940, 45757380, 36717114, 45757381, 441788, 4333885, 35610330, 36715556, 35610522, 4270602, 37109025, 42535207, 4084817, 4080770, 36716153, 4291602, 4306683, 4300215, 4081909, 4142828, 4185025, 4182586, 4200132, 3657814, 3657815, 4080331, 760906, 760907, 36685421, 44810378, 4302049, 36715819, 4345474, 4145196, 4028324, 137785, 140641, 40490394, 40491348, 40490302, 40489357, 4291600, 4294439, 4300214, 4294440, 4291603, 4130346, 36716496, 4296065, 3656108, 4111926, 4289145	
Hyperlipidemia	432867	
Hypertension	316866, 4322024, 42709887	4167493
Hyperthyroidism	140673	
Ischemic heart disease	4185932	
Myocardial infarction	4329847	314666
NAFLD	4026131, 40484532	
Obesity	4060985, 4256640, 45766204, 433736, 4176962, 4081038	
Osteoarthritis	80180	
Osteoporosis	80502	
Pancreatitis	4192640	
Peripheral vascular disease	321052	
Pneumonia	4050869, 255848	45770911, 4001167, 4049965, 36712839, 252552
Psoriasis	140168	
Pulmonary embolism	440417	
Renal impairment	4030518	

Rheumatoid arthritis	80809	
Stroke	42535426, 4048784, 4045735, 4031045, 761110, 372924, 4110189, 443454, 762951, 765515, 43530683, 762933, 762937, 4111714, 4108356, 45772786, 4110190, 762935, 763015, 46273649, 35610084, 46270031, 762934, 43531607, 35610085, 46270381, 4110192, 45767658, 44782773, 46270380, 37110678, 37110679, 381316, 35609033, 4046362, 4131383, 4046237, 4119140, 4043731, 439847, 4141405, 37116473, 4144154, 4111709, 4077086, 4046359, 4319146, 4043732, 4146185, 36717605, 43530727, 4148906, 43530728, 432923, 4108952, 4111708, 4142739, 4046358, 36684840	
Type 1 Diabetes	36715571, 45769891, 37016767, 45763585, 4128019, 4225656, 45773688, 45773576, 45771075, 45769902, 45769903, 45769837, 35626765, 45769832, 45757674, 435216, 42538169, 42535539, 377821, 37016353, 42689695, 45769904, 43531565, 4221344, 4223303, 37017429, 765533, 37016348, 45757432, 443592, 201531, 42535540, 45757393, 45771067, 45769876, 4228112, 45757362, 3046418, 4047906, 4102018, 45757073, 439770, 4224254, 4143857, 35626069, 45757535, 37016179, 43530660, 37016180, 4225055, 45769829, 45769830, 37312218, 45768456, 45763583, 45769834, 36713094, 318712, 37018566, 4222687, 4222553, 37017431, 4063042, 43531008, 43531009, 45763584, 45757604, 200687, 45757266, 4227210, 45770986, 45771533, 45773567, 45769833, 765373, 46269764, 4143689, 45769873, 201254, 40484648, 40484649, 4152858, 443412, 4099214, 45766051, 45757507, 45769892, 37312201, 45770902, 37312200, 45757074, 4224709, 765650.	
Type 2 Diabetes	4321756, 36717156, 43531588, 45769888, 4196141, 37016768, 609103, 609106, 609114, 609117,	

	602345, 45763582, 40483315, 4221495, 43531578, 43531559, 45769901, 43531566, 43531653, 43531577, 43531562, 37309630, 45769894, 43531616, 45757474, 36684827, 37018912, 443732, 43531597, 443733, 376065, 43531564, 45757280, 45769906, 4177050, 4223463, 43530690, 4222876, 37018728, 45772019, 604741, 37016349, 45770880, 201530, 4215719, 45757392, 45771064, 45757447, 45757446, 45757445, 45757444, 45757363, 45772060, 36714116, 608884, 45769875, 4130162, 45757075, 765375, 45771072, 443734, 4228443, 4140466, 45770830, 35626070, 45769905, 45757435, 609099, 609101, 43531651, 45770881, 609104, 609105, 4222415, 37162626, 45769828, 760989, 761063, 43531563, 45757450, 37312203, 37312202, 45770883, 37016354, 43530656, 609096, 609095, 45769836, 443729, 43530689, 45757278, 4221487, 4223739, 37017432, 3192767, 3191208, 3194332, 4063043, 43530685, 609116, 609119, 45770831, 45757499, 443731, 45770928, 4226121, 45769872, 45769835, 761053, 609109, 609112, 36712670, 46274058, 4142579, 45770832, 45773064, 201826, 45757508, 4230254, 4304377, 40485020, 4193704, 4200875, 4099651, 45769890, 37312205, 36712686, 45757277, 37312204, 36712687, 45757449, 43531608, 4099216, 761062	
Ulcerative colitis	81893	
UTI	81902	
Venous thromboembolism	762047, 762148, 761444, 35616028, 35615035, 761416, 35615031, 43531681, 35616027, 35615034, 761415, 35615030, 44782746, 44782751, 762008, 760875, 765155, 762017, 762417, 762020, 765546, 762004, 44782742, 44782747, 762015, 765541, 44782748, 44782752, 762009, 760876, 765540,	

	765922, 762418, 765537, 44782767, 46270071, 762022, 44782743, 762021, 762010, 760877, 762013, 762018, 762419, 762005, 44782745, 44782744, 762026, 765156, 44782421, 764016, 44782766, 4120091, 45768439, 45768888, 762048, 45757410, 762049, 36712892, 44782762, 37109253, 40478951, 4042396, 4046884, 4133004, 4181315, 45773536, 763942, 761980, 443537, 4133975, 40480555, 4322565, 763941, 761928, 4207899, 4028057, 435565, 40481089, 4309039, 4119760, 762808, 40480461, 4124856, 4096099, 440738, 4281689, 4284538, 4309333, 4108681, 46285905, 440417, 37109911, 37016922, 43530605, 254662, 4253796, 4121618, 4119610, 46271900, 4236271, 36713113, 35615055, 4033521, 4119607, 4055089, 4327889, 320741, 439838, 4230403, 4069561, 761831, 761830, 761808, 761832, 761809, 4221821, 440750, 4176614, 761821, 761819, 444097, 761820, 761818, 4110339, 4111868, 4110343, 439314, 4109877, 4112171, 4112172, 4250765, 42538533, 44811347, 765049, 4317289, 4203836, 4175649, 4149782, 4153353, 46285904, 444247, 77310, 4189004	
Viral Hepatitis	4291005	

Appendix 3. Preliminary code list of medications for large-scale characterizaton

Medications	Concept id (Included)	Concept id (Excluded)
Agents acting on the renin-angiotensin system	21601782	
Antibacterials for systemic use	21602796	
Antidepressants	21604686	
Antiemetics	21600490	
Antiepileptics	21604389	
Anti-inflammatory and Antirheumatic products	21603933, 21602722	
Antineoplastics	21601387	
Antipsoriatics	21602028	

Antithrombotic agents	21600961	
Antivirals for HCV	1501761	
Antivirals for HIV	21603180	
Beta blocking agents	21601664	
Calcium channel blockers	21601744	
Diuretics	21601461	
Drugs for acid related disorders	21600046	
Drugs used in addictive disorders	21604816	
Drugs for obstructive airway diseases	21603248	
Drugs for diabetes	21600713, 21600744	
Hormonal Contraceptives	21602472	
Immunosuppressants	21603891	
Iron preparations	21601078	21601119
Lipid modifying agents	21601853	
Opioids	21604254	
Psycholeptics	21604489	
Psychostimulants	21604752	