Data Sources for Studying Genetic Factors of Chronic Obstructive Pulmonary Disease (COPD)

AJ Kapasi[1], SA Kamani[1], **JK Jones[1†]**

[1]DGI, LLC (nonprofit), Fairfax, VA, USA E-mail: info@bridgetodata.org

†JK Jones is deceased (August 4, 2020)

BACKGROUND

- ✧ Genetic epidemiologists research genetic factors associated with the risk of disease.
- ✧ Genome-wise association studies (GWAS) are a powerful first step in detecting genetic loci associated with diseases such as COPD. Findings must be confirmed with high-powered, large-scale gene-disease association studies.
- ✧ A critical step is obtaining robust data samples that allow sufficient power to measure genetic associations with disease. A researcher must know which healthcare databases (DBs) are available, what data are captured, and their strengths/limitations.
- ✧ B.R.I.D.G.E. TO DATA® (B.R.I.D.G.E., www.bridgetodata.org), an international compendium of population healthcare DB profiles, is a unique genetic epidemiology resource.

OBJECTIVES

To show how genetic epidemiologists identify and compare genetic healthcare DBs through a case study relevant to drug developers.

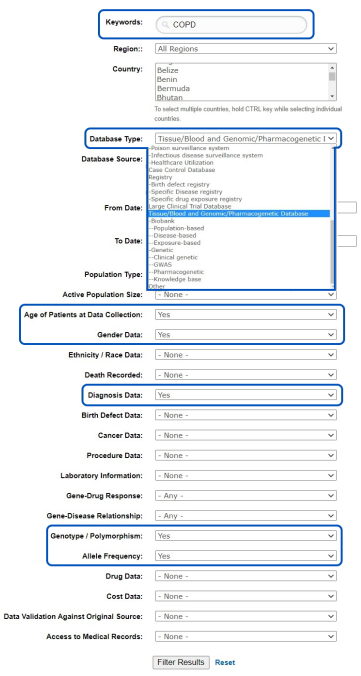
METHODS

CONTEXT – A researcher has identified new genetic variants associated with reduced lung function through GWAS. The drug developer must now confirm potential COPD targets in one or more genetic healthcare DBs.

SEARCH STAGE – A search was conducted in B.R.I.D.G.E. (324 DB profiles worldwide as of September 8, 2020) to identify healthcare DBs that contain a population of COPD patients and collect genetic data (**Figure 1**):

Keyword = COPD
DB Type = Tissue/Blood/Genetic/Pharmacogenetic
Age = Yes
Gender = Yes
Diagnosis = Yes
Genotype/Polymorphism = Yes
Allele Frequency = Yes

Figure 1. B.R.I.D.G.E. TO DATA® Search Page



REVIEW STAGE – Search results were manually reviewed and selected for comparing data attributes (**Table 1**).

FEASIBILITY STAGE – Feasibility of a pooled analysis to achieve sufficient power was assessed.

Table 1. Examples of Data Fields Used in Profiles (by Category)

Category	Data Fields
Summary	Database description, Database source, Years covered, Population type, Date of last update
Population Dynamics	Population size, Sample weights – Extrapolation factors
Demographic Data	Age, Gender, Date of birth, Death recorded, Other demographic data
Physician & Practitioner Info	Physician ID & Specialty, Pharmacy ID
Diagnoses/Signs & Symptoms	Diagnosis data, Diagnoses coded (coding systems), Max. number of codes, Physical exam findings, Environmental exposures, Behavioral data elements
Procedures	Procedure data, Procedures coded (coding systems), Laboratory information
Drug Information	Drug data, Drug dosage, Drug coding system(s), Additional drug information
Biobanks	Human specimen, Biomarkers, Pre-diagnostic/Post-treatment sample collection, DNA/RNA isolation, Family/Medical history
Genetic-PGx Data	Source of genetic data, Method of imputing/variant filtering, Haplotypes, VNTR, SNPs, Genotype/Polymorphism, Allele frequency, Linkage disequilibrium
Economic Data	Type of cost data (if applicable)
Validation & Linkage	Data validation, Access to medical records, Linkage to other databases
Administrative Data	Database contact data, Database usage restrictions, References of studies using/describing the database

RESULTS

SEARCH STAGE – The search yielded 2 relevant hits (>75% match): *Women's Health Initiative* (WHI; USA) and *UK Biobank*. Manual review of lower relevancy-ranked profiles (≤75%) identified additional DBs with the potential to provide genetic data linked to COPD outcomes: *Maccabi Health Services (MHS) COPD Registry* (Israel), *BIG3* (Sweden), and *PHARMO* (Netherlands).

REVIEW STAGE – *WHI* and *UK Biobank* contain genetic data linked to medical, demographic and lifestyle (e.g., smoking) information (**Table 2**). Both DBs consist of clinical trial participants, but *WHI* is limited to postmenopausal women. *UK Biobank* reports small nucleotide polymorphisms (SNPs), insertions/deletions, allele frequencies, and supports public access. Some SNPs in *WHI* are publicly available (dbGaP), while other data are restricted to WHI investigators.

MHS COPD Registry is a subset of MHS Clinical DB; biospecimens from adults are being collected (Tipa Biobank) for eventual genetic analysis. *BIG3* and *PHARMO* allow genotyping of existing or prospectively obtained blood samples, respectively, with special authorization.

FEASIBILITY STAGE – Combining *WHI* and *UK Biobank* study cohorts is not ideal given the major differences in study aims, sampling, and demographics.

Table 2. Excerpt from B.R.I.D.G.E. TO DATA® Comparing Data Elements in 3 Selected Databases with Genetic Data

FIELD NAMES	Women's Health Initiative (WHI) (USA)	UK Biobank (United Kingdom)	BIG3 (Sweden)
Region	The WHI Clinical Trial and Observational Study were both conducted at 40 clinical centers in 24 states and the District of Columbia	Includes volunteers across UK. While UK Biobank is not nationally representative and subject to "healthy volunteer bias", valid assessment of exposure-disease relationships is still widely generalizable	Southern (Skåne) region
Database Type	•Longitudinal Population Database (Outpatient Drug & Diagnosis Data) •Large Clinical Trial Database •Genetic (GWAS) Database •Observational Study The randomized controlled Clinical Trial enrolled 68,132 women into trials testing 3 prevention strategies (Hormone therapy, dietary modification, Calcium/Vitamin D). Observational Study tracks medical events & health habits of 93,676 women. Recruitment for observational study was completed in 1998; participants have been followed since. WHI holds a large repository of biological specimens available for ancillary study investigations. Purpose of biorepository is to gain insights into women's susceptibility to certain diseases, which treatments are best for which women, and how to individually tailor preventive care.	•Longitudinal Population Database (Drug & Diagnosis Data) •Large Clinical Trial Database •Biobank (Population-based) •Genetic (GWAS) Database UK Biobank is a large, population-based prospective study. It recruited 502,642 40-69 year-olds across UK (2006-2010). Participants have undergone measures, provided biological samples for future analysis, detailed information about themselves, and agreed to have their health followed. Purpose of this biobank: Conduct detailed investigations of genetic & non-genetic determinants of diseases of middle & old age. With consent, participant data are being linked to their health-related records, so baseline information can be used in conjunction with the information about health conditions that develop. Genome-wide genotyping data are available for all participants.	•Longitudinal Population Database (Disease-specific Diagnosis Data) •Registry •Clinical Genetic Database Open prospective longitudinal cohort study, where registries are used in the initial selection process. Blood samples are analyzed both, directly for a number of markers, and also stored in a biobank for later analysis. BIG3 aims to find out which individuals are at greater risk of developing COPD, cardiovascular disease, and/or lung cancer, and why some individuals seem to be protected.
Database Source	-Case Report Forms (Self-administered forms) -Survey Data (Interviews) -Clinical measurements -Biospecimens collection (blood, urine)	-Case Report Forms (Self-completed touch-screen questionnaire) -Survey Data (computer-assisted interview, web-based questionnaires) -Physical, functional measures & activity data (wearable monitors) -Biospecimens collection (blood, urine, saliva, genetic analyses) -Imaging data	-Survey Data -Registry -Laboratory data -Blood samples
Years Covered	1993 - Present WHI (original): 1993 - 2005 WHI Extension Study: 2005 - 2010 (1st); 2010 - 2020 (2nd)	2006 - Present (Recruitment from 2006 through 2010)	2013 - Present (The estimated completion date of data collection is December 2020)
Population Type	-Outpatient/Non-Institutionalized -Clinical Trial Participants (Women in WHI were aged 50-79 & postmenopausal at time of enrollment. Exclusion criteria varied for each study in the Clinical Trial and Observational Study component.)	-General Population (Participants were assessed in 22 assessment centers throughout UK, covering a variety of different settings to provide socioeconomic and ethnic heterogeneity and urban-rural mix.)	-General Population (Initial screening, ~100,000 people in Scania aged 45-75. Then ~10,000 individuals from the respondents are summoned to undergo examinations. The goal is to try to get ~25% of smokers in the cohort.)
Database / Final Population Size	<200,000 (161,808 is the final total population size of the WHI study; however, with each Extension Study or Ancillary Study, the population size decreases)	0.5 - 1 Million (A total of 502,642 participants were recruited; follow-up data collection and analyses are still ongoing)	<200,000 (Data collection is ongoing; estimated completion date for BIG3 data collection is December 2020. As of Sep. 2014, number of respondents to questionnaire was ~3,000.)
Age (%) of Patients at Data Collection	Yes <18 years = 0% >60 years = 66.9%	Yes <18 years = 0% >65 years = 18.4%	Yes <18 years = 0 participants >65 years = 2,056 participants
Gender (%) Data	Yes (Females = 100%)	Yes (Females = 54.4%)	Yes
Ethnicity / Race Data	Yes	Yes	No
Other Demographic Data	Yes (Education, Employment, Marital status, Income, Social support, etc.)	Yes (110 fields: Education, Employment, Household, Early Life, Indices of Multiple Deprivation, etc.)	Yes (Respondents complete a questionnaire about lifestyle, living environment, and tobacco habit)
Diagnosis Data	Yes: Cardiovascular, Cancer, Fractures, Other (including COPD), Reproductive history, QOL, Cognitive assessments, etc. -Diagnoses from hospital discharge records: ICD-9-CM -Cancer diagnoses: ICD-O-2 and SEER EOD during adjudication	Yes: Diagnosis data are captured in the main and secondary diagnosis data fields as well as data fields for cancer, death, family history, medical history, psychosocial, cognitive, physical measures, and sex-specific factors using ICD-9 and ICD-10	Yes: COPD, Cardiovascular diseases, and Lung cancer. ICD-10 codes are available on a sub-project basis.
Environmental Exposures	Yes (e.g., Insecticides, Pets, Sunlight, Secondhand smoke)	Yes (e.g., Occupational hazards, Sun exposure, Secondhand smoke)	Yes (Lifestyle, Living environment, and Tobacco habits)
Behavioral Data Elements	Yes: Smoking, Alcohol, Thoughts and feelings, Daily life items, Diet/Nutrition, Sleep patterns, etc.	Yes: Smoking (~30 data fields), Alcohol, Diet, Physical activity, Sexual factors, Sleep, Screen use, Driving, etc.	Yes: Smoking, Alcohol
Procedure & Laboratory Data	Yes: Self-reported and ICD-9-CM procedure codes from clinical center physicians who obtain patient hospital discharge records	Yes: Data on operations (OPCS) and laboratory testing are captured	Yes: Respondents are summoned to undergo examinations
Drug Data	Yes: Prescription & OTC (List of current medications & supplements used, and some medications/supplements have their own questionnaire, e.g., hormone replacement therapy)	Yes: Common prescription, OTC, Illicit drugs, Supplements, Herbals, Flu vaccines, etc.	No
Biobank Type	Clinical trial biorepository	Population-based prospective study with biospecimens	Disease-based study with biobanking of blood samples
Biomarkers	Yes: Over 600 biomarkers have been tested	Yes: Established disease risk factors (e.g., lipids for vascular disease), diagnostic measures (e.g., HbA1c for diabetes), require further assessment (e.g., biomarkers for renal and liver function), and 820,967 genetic markers of disease included on the UK Biobank Axiom Array	Currently not available
Type of Genetic Database and Genetic Testing	GWAS database: GWAS has been performed through many WHI ancillary studies with different platforms and different outcomes/exposures of interest, but GWAS data from about 30,000 WHI participants were imputed into 1,000 Genomes data. The harmonization/imputation effort involves 6 different GWAS studies: 1000 Genomes Project reference panel (1092 samples; v2.20101123 for GECCO; v3.20101123 for Hip Fracture, SHARE, GARNET, WHIMS + MOPMAP). The Harmonized and Imputed GWAS data are available at dbGaP (phs000746), but not all directly genotyped data have been submitted. Genetic and phenotypic data from WHI sub-studies are routinely submitted to the database of genotypes and phenotypes (dbGaP). Individual-level genetic and phenotype data are available for 64,291 study subjects.	GWAS database: Genome-wide genotyping data are available for all 500,000 participants in the UK Biobank cohort. Genotyping was performed using the Affymetrix UK BiLEVE Axiom array on an initial 50,000 participants; the remaining 450,000 participants were genotyped using the Affymetrix UK Biobank Axiom® array that genotyped ~850,000 variants. The two arrays are extremely similar (with >95% common content)...Other genetic sequencing projects are underway (e.g., Exome sequencing, Whole genome sequencing).	Not applicable
Gene-Drug Response	Yes: Drug data can be linked by subject ID with genetic data	Yes: Pharmacogenetics as well as ADME content consists of markers for genetic variants of related genes listed in the PGKB	Not applicable
Gene-Disease Relationship	Yes: The current release of outcomes data includes centrally verified, locally verified and self-reported outcomes collected through the first WHI Extension Study for clinical trial, observational study, and Calcium and Vitamin D study. The outcomes data sets include all WHI participants, and the first occurrence of outcomes since the beginning of WHI. Outcomes of interest are adjudicated or laboratory-tested.	Yes: UK Biobank Axiom® Array is designed using imputation-aware SNP selection. This array provides optimized content modules for genome-wide association studies (GWAS) of common and low-frequency variants, biological function, and human disease in populations of European and British ancestry. The comprehensive coverage also includes rare coding variants, pharmacogenomics markers, copy number regions, HLA, inflammation, and eQTL variants.	Currently not available
Variant Type	SNPs	SNPs, Insertions, Deletions, Large structural variants	Not applicable
Allele Frequency	Yes: Nucleotide substitution (e.g., C/T)	Yes: Information scores and minor allele frequency data for the imputed genotypes are available	Not applicable
Data Sharing: Genetic Data	Yes: Access to genetic data is available - only epidemiologic research	Yes: UK Biobank intends to make available a set of all (or majority of) GWAS results available through the European Genome Archive	Not applicable
Linkage to Other Databases	Yes: The WHI Central Coordinating Center (CCC) developed a Virtual Data Enclave (VDE) that allows Investigators to securely access participant Medical history and medical services use, Residential history, Vital statistics, etc. Genetic and phenotypic data are routinely submitted to the public dbGaP. In 2010, the WHI SNP Health Association Resource (SHARe) dataset was released on dbGaP. A subset of the WHI Clinical Trial and Observational Study data is also available through public BioLINCC. Other study-specific data linkages have been performed (e.g., linkage to the Census 2000 data for the AS464 study).	Yes: Identifiable data (such as name, date of birth, NHS number) are collected for each participant, and acts as a unique identifier for linkage purposes. Genetic data linkages to health outcomes data sources include cancer and death registers, hospital discharge diagnosis data, general practitioner data, and other medical (e.g., prescriptions, pathology reports, imaging reports, screenings) and health-related data (e.g., employment, benefits, socio-economic records).	Yes: Data will be periodically linked to the regional EHR (i.e., linking information regarding if the participant's health status is changing over time, and if so, what the specific health status changes are)
Database Usage Restrictions	Public & Private Access: Access to WHI Datasets is governed by study policy. Current policy requires that in order to gain access to WHI Datasets, you must be a current/ancillary study/former PI or lead author on an approved paper. De-identified genotyping data can be requested at the BioLINCC website.	Public Access: Data are available to all bona fide researchers for all types of health-related research that is in the public interest, without preferential or exclusive access for any person. Access to biological samples that are limited and depletable will be carefully controlled and coordinated.	Private: Data are not publicly available online; researchers must contact the database manager to request access to BIG3 data

CONCLUSIONS

- ✓ This case study demonstrates how B.R.I.D.G.E. supports the search, review, and feasibility stages of the DB selection process for gene-disease association studies.
- ✓ **B.R.I.D.G.E. was successfully used to identify genetic healthcare DBs**, compare attributes, and assess potential for pooling cohorts.
- ✓ **Limitations:** This analysis was a **limited sampling** using DBs currently profiled within B.R.I.D.G.E. TO DATA®. **More profiles of data sources are continually being added** to B.R.I.D.G.E. Future analyses may provide a better comparison.
- ✓ As B.R.I.D.G.E. grows, it may be **a tool for standardizing healthcare DBs**, including those with genetic data.

This study was self-funded by the nonprofit organization, DGI, LLC