**Introduction:**
Clinical Trial Protocols define the target patient population through various inclusion and exclusion criteria. These criteria may be well justified by safety, scientific, regulatory, or operational reasons, but very often limit the number of available patients by a significant amount. TriNetX is a federated network, providing real world data based on Electronic Health Records (EHR), which can be used to test Clinical Trial Protocols for feasibility, especially evaluating the impact of individual in- or exclusion criteria on patient availability (Fig 1). In a second step, the patient cohort as defined by the criteria can then be located at member institutions and can be re-identified there for potential enrollment into the trial.

Although the network is continuously growing and has a diverse portfolio of provider members, it needs to be shown how well the network population represents the general population.

**Objective:**
To evaluate whether a federated data network of health care providers supplying aggregate counts on electronic health records (EHR) is currently representing a real world disease landscape so it can be used for development and feasibility testing of clinical trial protocols in various therapeutic areas.

**Method:**
ICD-9 diagnoses provided by a network of 7 health care providers in the US, consisting of 11 million patient million, were analyzed. TriNetX cohort analyzes were compared to select published epidemiologic data in therapeutic areas of interest for the pharmaceutical industry with special focus on:

- Frequency of diagnoses
- Demographics, age distribution, including smoking status
- Correlation of various risk factors and complications

**Results:**

**Most frequent diagnoses:**
As of January 2016, 11,973,320 patients in the age of 1 month to 88 years. The information is updated approximately every two to four weeks. The three most frequent individually coded diagnoses were essential hypertension, hyperlipidemia and diabetes mellitus in 979,822 (8.6%), 883,441 (7.8%) and 429,999 (3.8%) of all patients, respectively. These diagnoses were identical to the three most common conditions as the US in the four published literature. Other frequent diagnoses were acute upper respiratory infection 548,073 (4.8%), asthma 401,074 (3.5%), anxiety 345,527 (3.0%), and lumbago 322,890 (2.8%).

**Smoking status:**
ICD code 305.1 for tobacco use was provided for 179,999 patients, 96,731 (53.7%) male and 83,268 (46.3%), comparable to the gender distribution published by the American Heart Association (55.5% and 44.5%).

**Diabetes and cardiovascular risk:**
Of the 266,592 patients in the network with ICD-9 240-250 for diabetes mellitus, a total of 1741 (2.2%) patients had at least one episode of hypoglycemia, compared to 2.1% in the ADVANCE study. 11144 (4.2%) of the patients with diabetes mellitus had myocardial infarction which constitutes a 12-fold higher frequency compared to 27251 (0.3%) patients in the TriNetX population without diabetes.

**Diagnosis demographics:**
There were 47,588 patients with a diagnosis of prostate cancer in the network. 46,887 (98.5%) were aged 50 years and older, 60% white and 26% black or African American. While the mean age was 73 years in the network, CDC’s National Program on Cancer Registries (NPCR) reports a general trend to a decreasing mean age (67.2 years).

**Conclusion:**
The network provides statistics on diagnoses, procedures, demographics, laboratory values, and medications. Since ICD coding is primarily done for billing, its representativity for other purposes was to be validated.

We compared network statistics with select epidemiologic publications (using literature as “gold standard”) and found vast consistencies:

- The most common diagnoses in the network were identical to the most common conditions diagnosed by physicians in the US, i.e. cardiovascular disease, obesity and diabetes.
- Gender distribution of smokers was similar to the data published by the American Heart Association.
- The percentage of patients with diabetes mellitus (DM) with an episode of hypoglycemia was close to the one observed in the ADVANCE study.
- 6.3% of the patients with DM had a diagnosis of MI, a 12-fold risk compared to the base population. Published data about diabetes-related complications in the United States shows a similar risk, i.e. 10-9-fold.
- The age distribution of patients with prostate cancer matched NPCR data. However, NPCR reports a decreasing mean age, lower than in our network.

**References:**
(2) Heart Disease and Stroke Statistics – 2015 Update: Circulation, 2015, 131, Table 3-1.