Why is it important?

You run a study in the OHDSI Network. You created perfect cohorts that precisely reflects the patients you want to study. But will you be able to find the concepts you used in the study datasets?

What did we do?

17 datasets from the US, Korea, Australia and Japan; EHR and claims. Rounded small counts to 100 and analyzed the distribution and granularity. For Figures 3-4 we took the concepts with their descendants and plotted their total counts X axis and number of datasets on Y axis. We scaled the data points depending on the counts of verbatim concepts and colored them based on the number of datasets that contain verbatim concepts.

*Some datasets contained limited variety of concepts (MIMIC3, NHIS Korea, HCUP) which reduced the number of overlapping concepts. We obtained only condition concepts from Ajou University dataset and condition, measurement and observation concepts from NHIS Korea dataset.
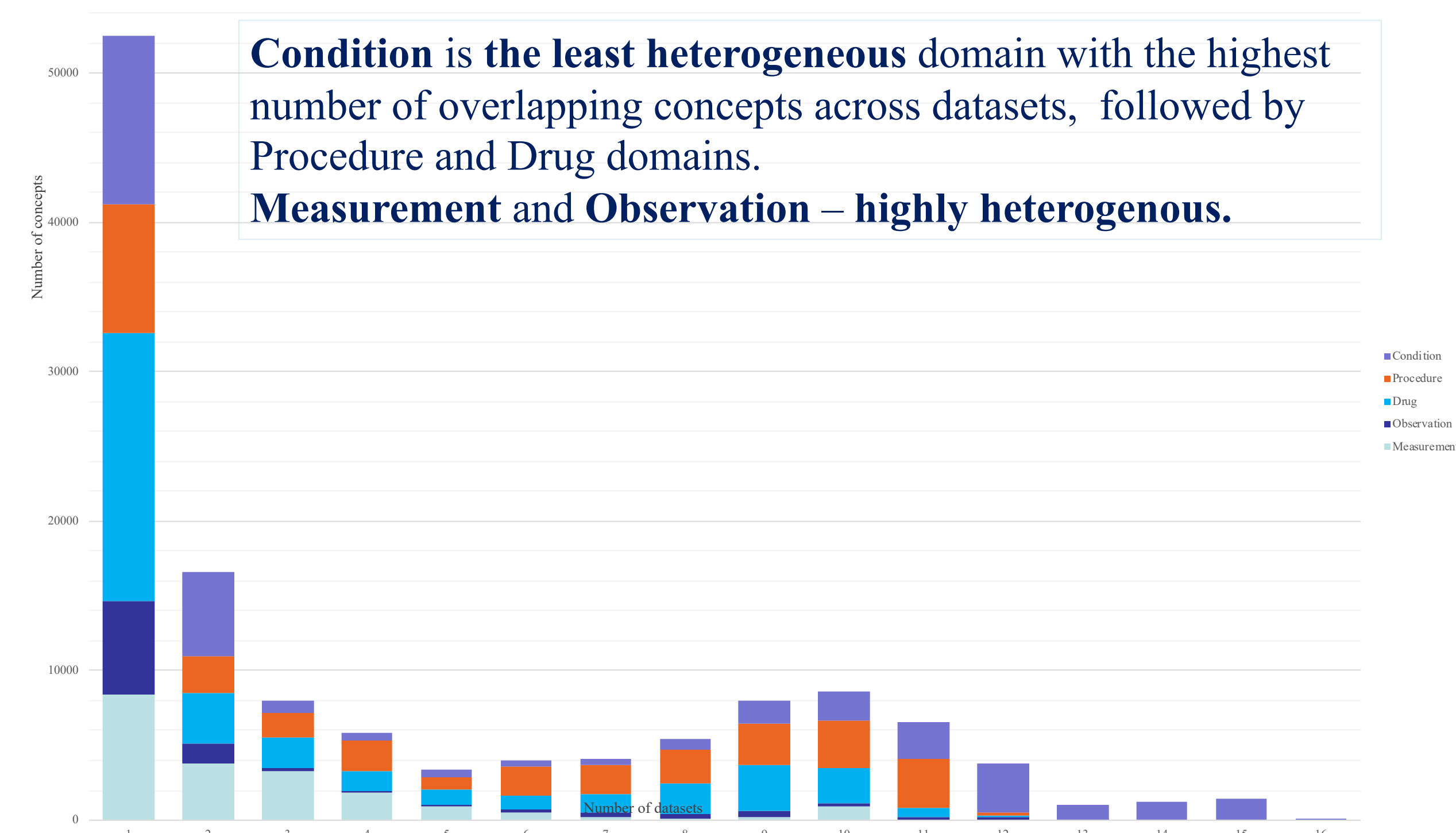
What did we find out?

Condition is the least heterogeneous domain with the highest number of overlapping concepts across datasets, followed by Procedure and Drug domains.
Measurement and Observation – highly heterogenous.

Figure 1. Distribution of the overlapping concepts across OHDSI network.

Surprise! Some of the highly prevalent concepts can be found only in a few datasets (Like 'Caffeine user' or 'Platelet distribution width [Ratio] in Blood').
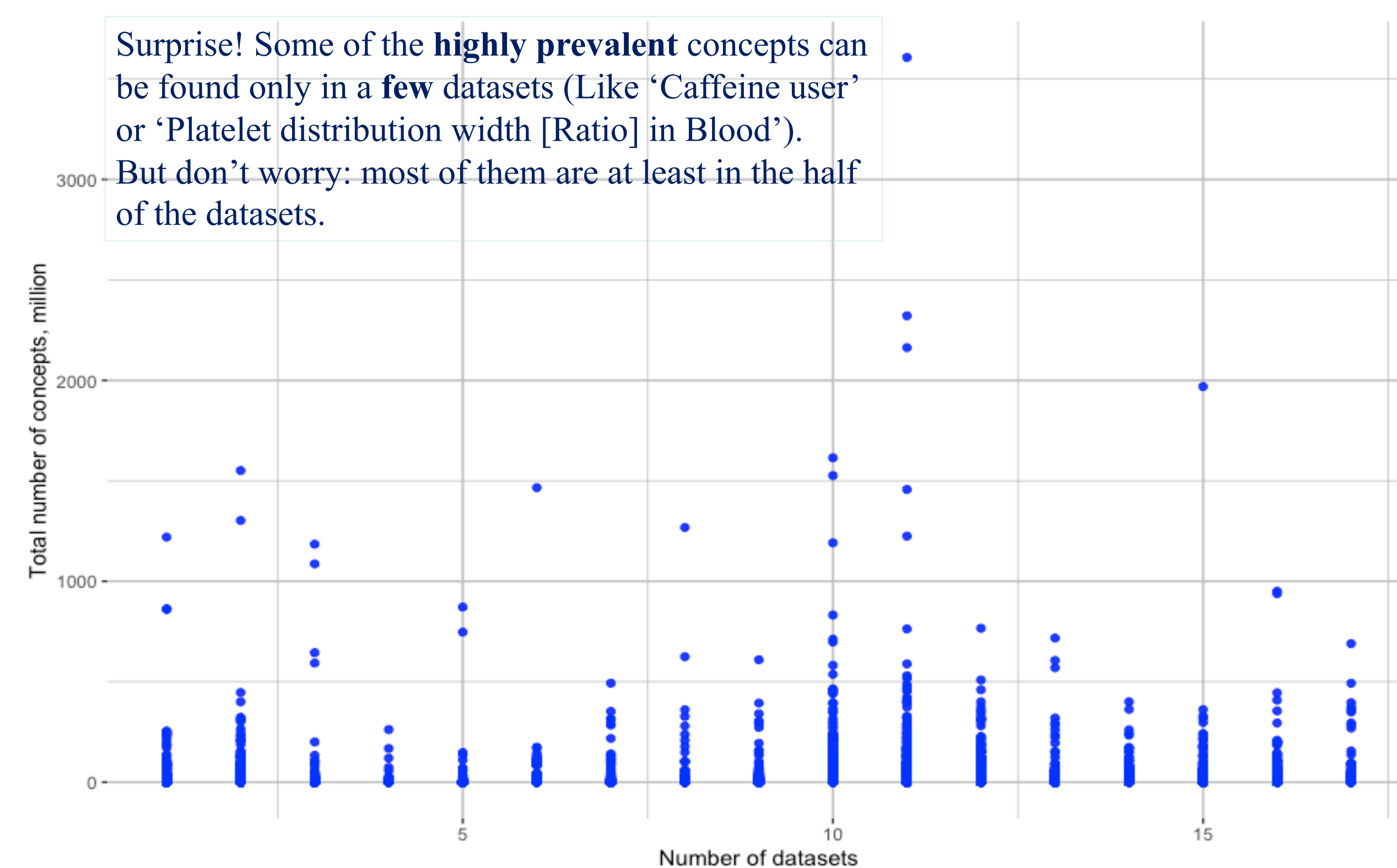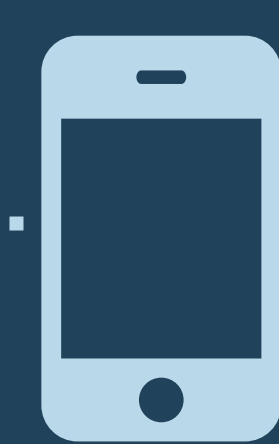But don't worry: most of them are at least in the half of the datasets.

Figure 2. Distribution of frequency of the concepts across OHDSI network

The majority of concepts don't appear in most databases, so all network studies require use of comprehensive concept sets when defining cohorts to represent the diversity of data across the community.

Take a picture to download the abstract

Concept Heterogeneity and Granularity in the OHDSI Network

Anna Ostropolets[1], Anthony Molinaro[2], Frank DeFalco[2], Jitendra Jonnagaddala[3], Siaw-Teng Liaw[3], Hokyun Jeon[4], Rae Woong Park[4,5], Matthew E. Spotnitz[1], Karthik Natarajan[1], George Argyriou[6], Kristin Kostka[6], Christian Reich[6], Chunhua Weng[1], Patrick Ryan[1,2]
[1] Columbia University, New York, USA; [2] Janssen Pharmaceuticals, New Jersey, USA; [3] University of New South Wales Sydney, Australia; [4] Ajou University Graduate School of Medicine, Suwon, Gyeonggi-do, Republic of Korea; [5] Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea, [6] IQVIA, Cambridge, MA, USA
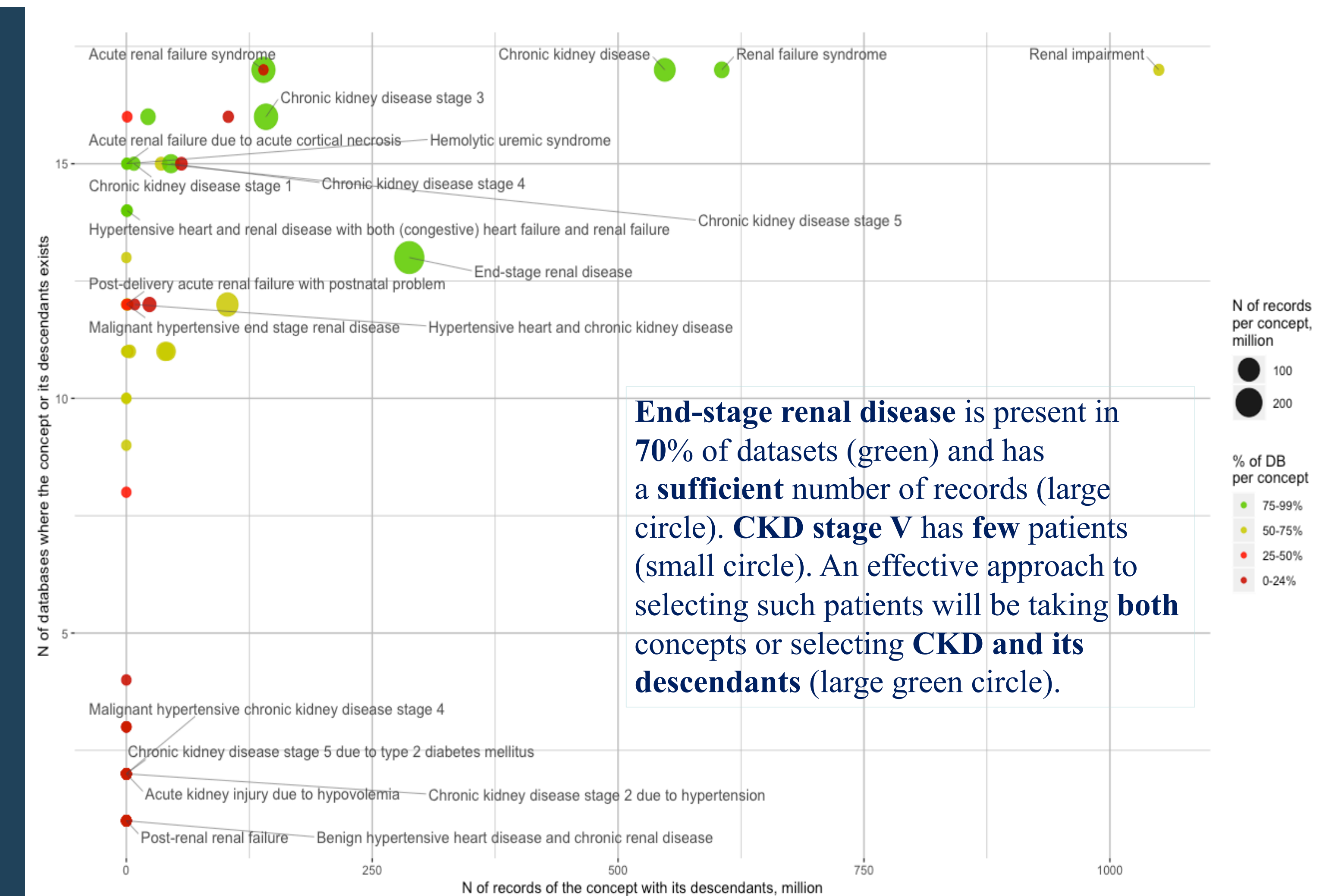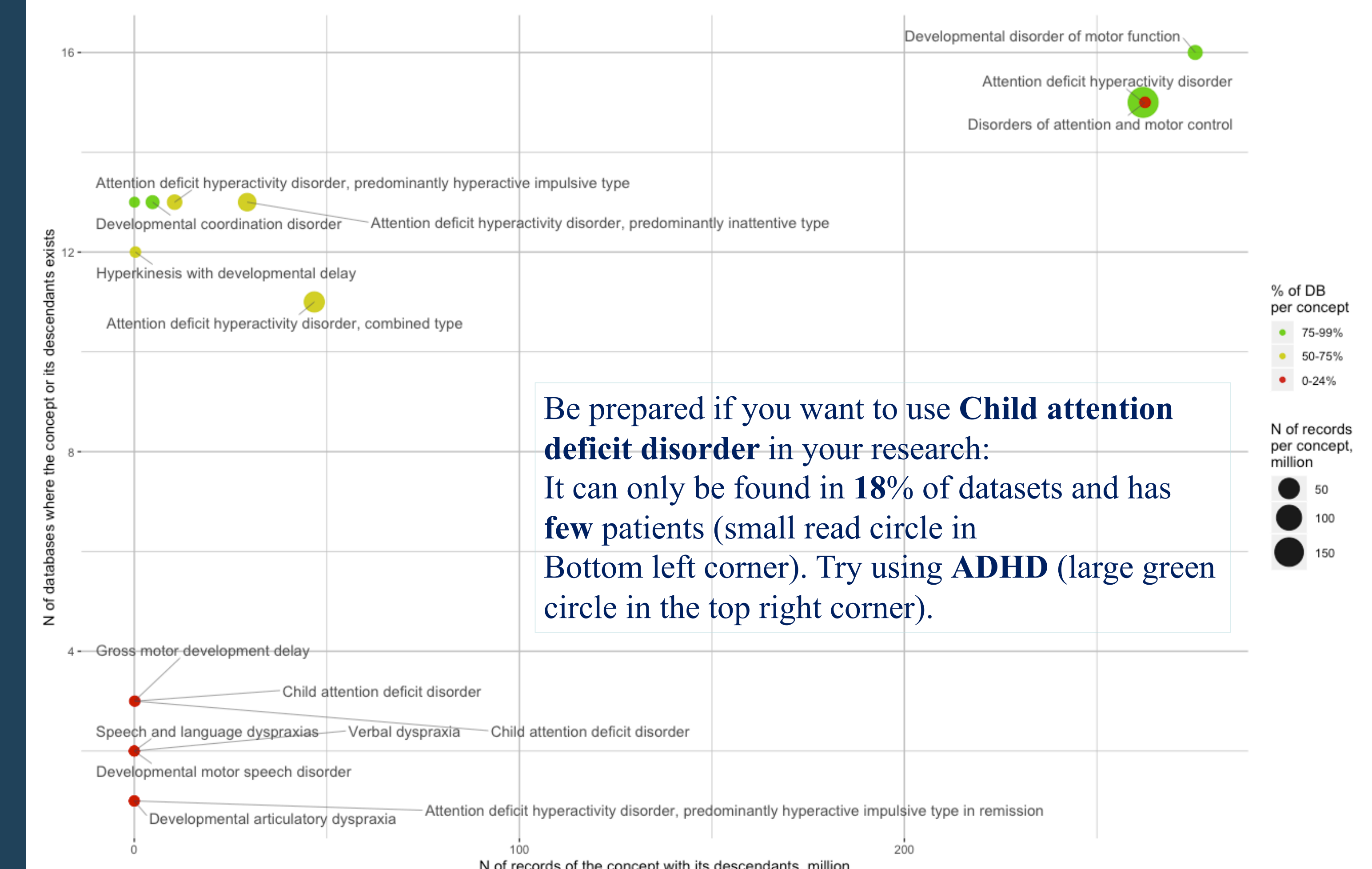
End-stage renal disease is present in 70% of datasets (green) and has a sufficient number of records (large circle). CKD stage V has few patients (small circle). An effective approach to selecting such patients will be taking both concepts or selecting CKD and its descendants (large green circle).

Figure 3. Renal Impairment

Be prepared if you want to use Child attention deficit disorder in your research:
It can only be found in 18% of datasets and has few patients (small read circle in Bottom left corner). Try using ADHD (large green circle in the top right corner).

Figure 4. Attention deficit disorder.

You cannot get away with using broad concept for UTI:
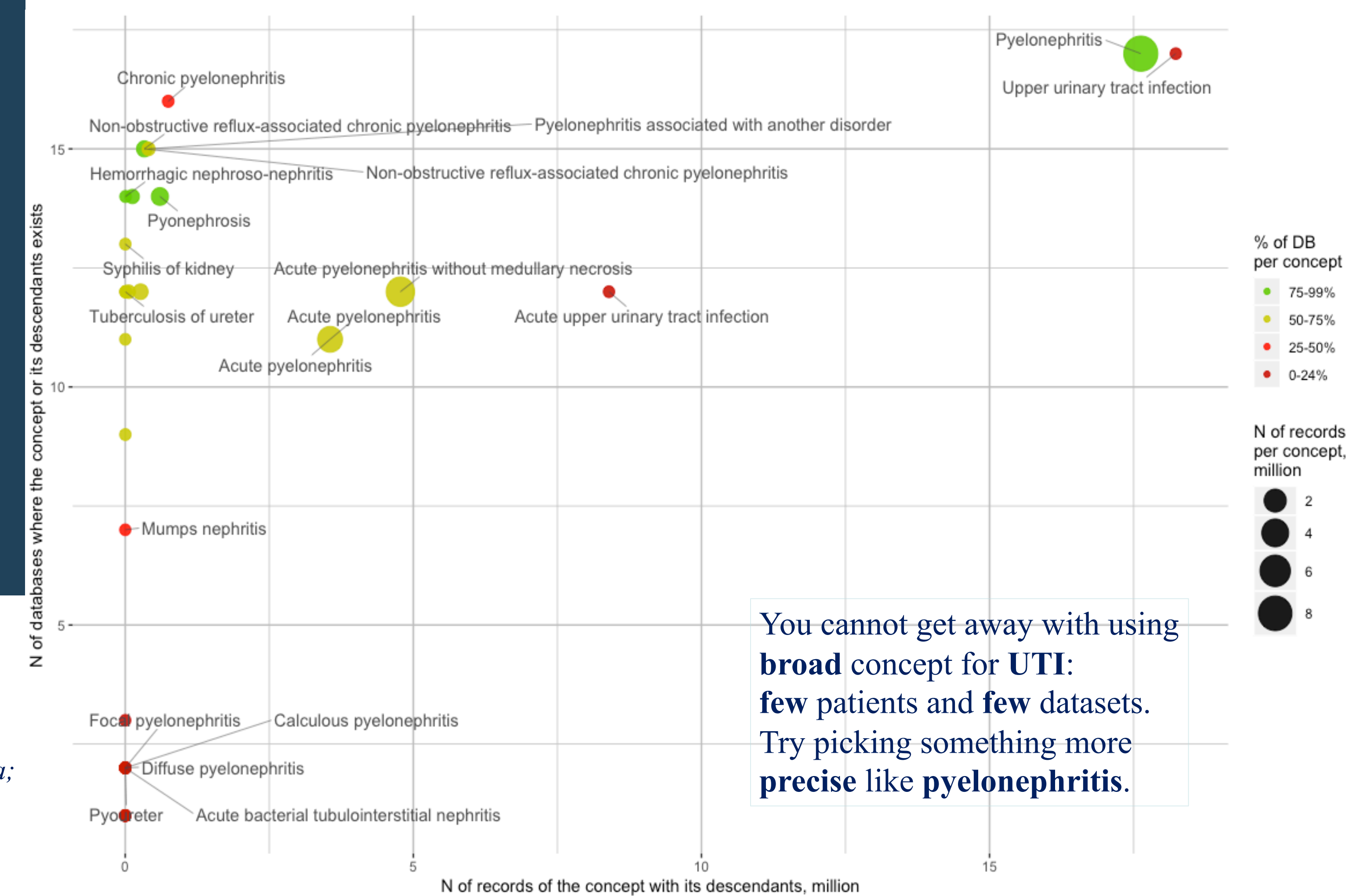few patients and few datasets.
Try picking something more precise like pyelonephritis.

Figure 5. Urinary tract infection.