STARR Data Synopsis

We operate STARR, a research data repository with 20 years of fully identified clinical data. STARR includes, but is not limited to, nightly clinical data, Epic Clarity, from both Stanford Health Care (adult hospital) and Stanford Children's health (aka Lucile Packard Children's Hospital). STARR contains not only fully identified and up-to-date (within the last 24-36 hours) Epic data from both SHC and Packard but also fully identified and current imaging data from Radiology as well as some historic clinical data from earlier EMRs that is not present in the current EMR systems

This invaluable data repository supports self-service tools for cohort identification (https://stan.md/2tjuLpl) and chart review (https://stan.md/2tntInf).

The data in STARR has many fascinating stories to tell. For example, patients come to Stanford from all over the country - indeed, all over the world - since for decades Stanford has been a world-class destination for specialty care.

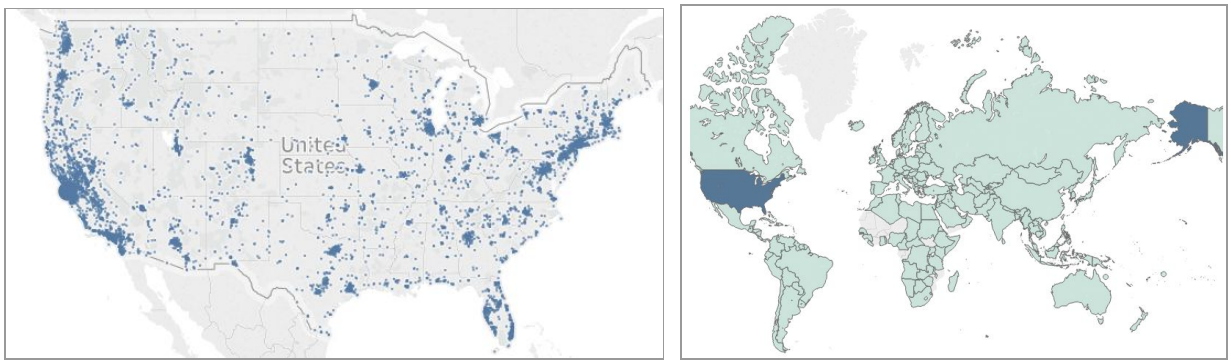

Figure 1: Heat map showing zip codes and countries of residence for patients in STARR

As another example, since all clinical encounters have a visit provider, and the provider record contains the provider's clinical speciality, it is possible to ask questions of the data regarding the frequency of visit for providers by speciality.

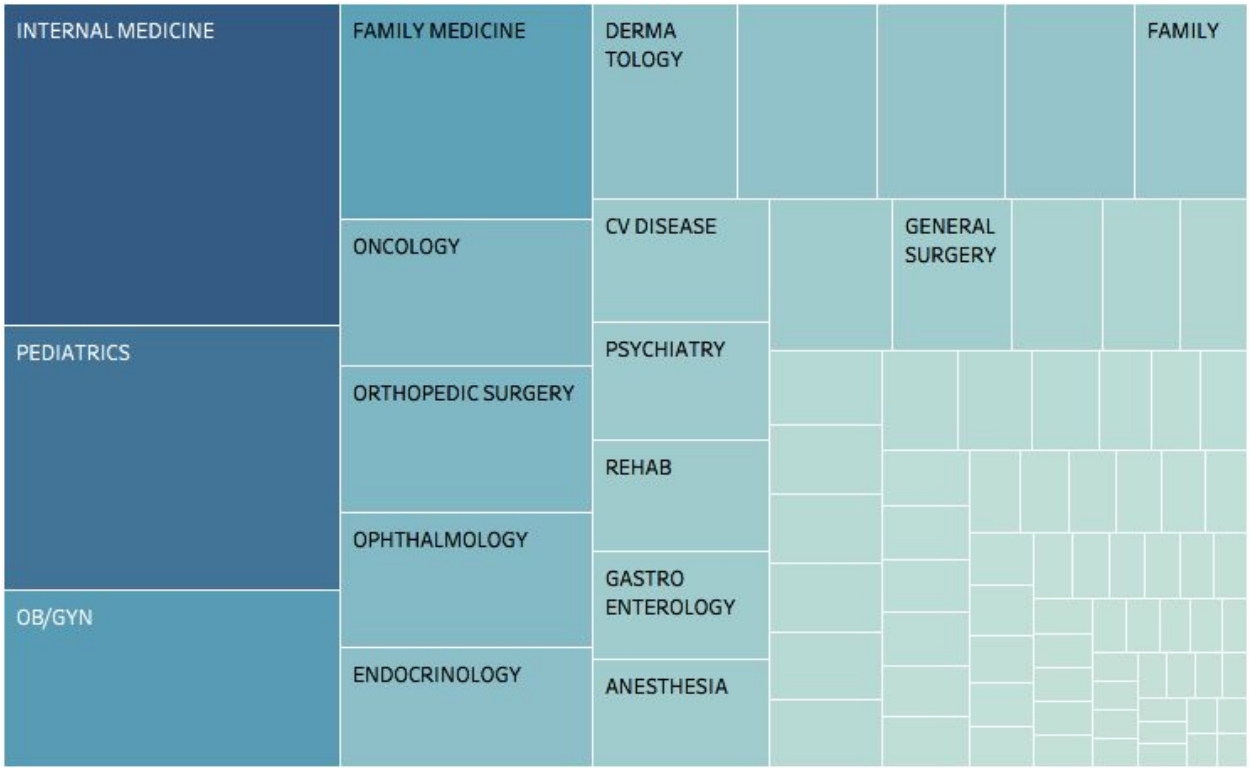

Figure 2: a frequency map of provider specialties drawn from the provider data in STARR

The number of patients seen at Stanford has been climbing since 2012 due to new outpatient clinics.

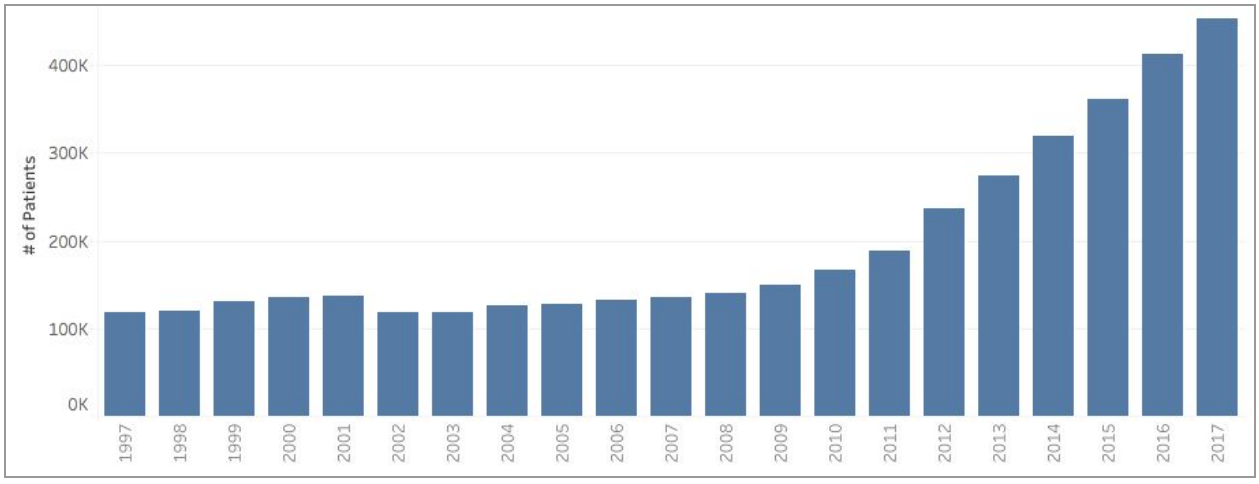

Figure 3: number of patients seen per year since 1997, drawn from STARR

And the sheer volume of data collected on these patients has been skyrocketing, from around 1.5M electronic records captured per year to over 11M per year in 2017.

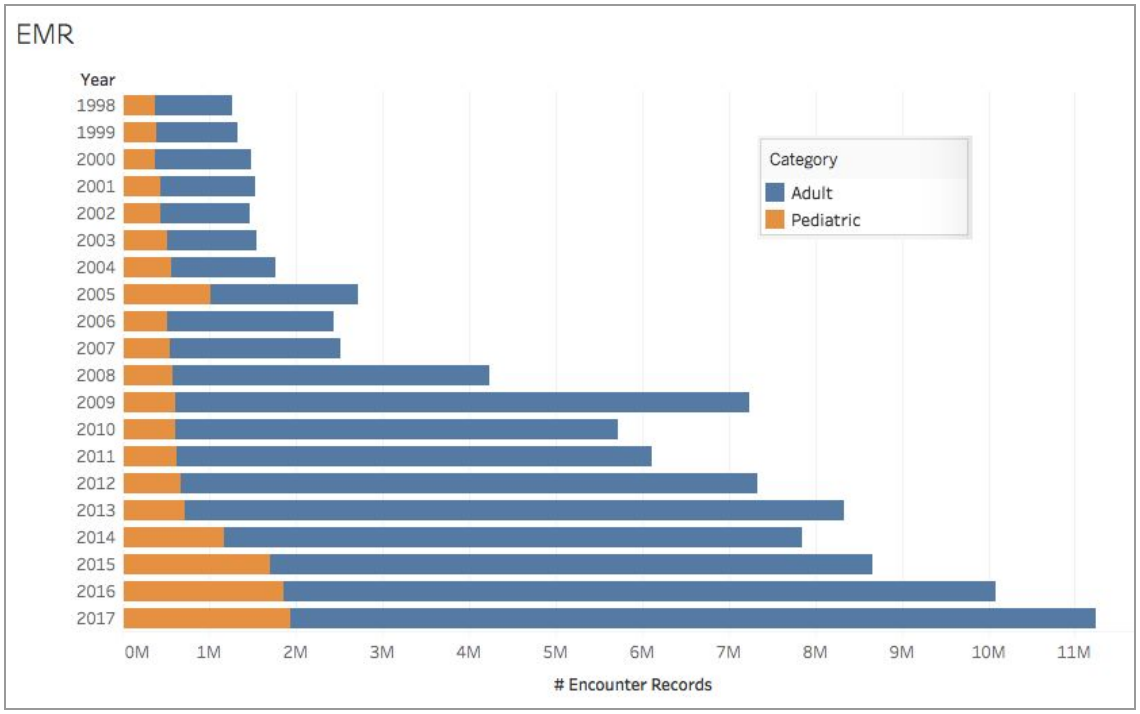

Figure 4: number of encounter records created per year since 1997, drawn from STARR. The spikes in electronic record generation correspond to EMR implementations: in 2005, Packard implemented Cerner, and in 2009 SHC implemented Epic.

The data exists in a relational database which can be queried to create custom datasets. STARR comprises some of the most frequently queried types of data: patient, provider, encounter, diagnoses, procedures, pharmacy orders and administration, and lab results.

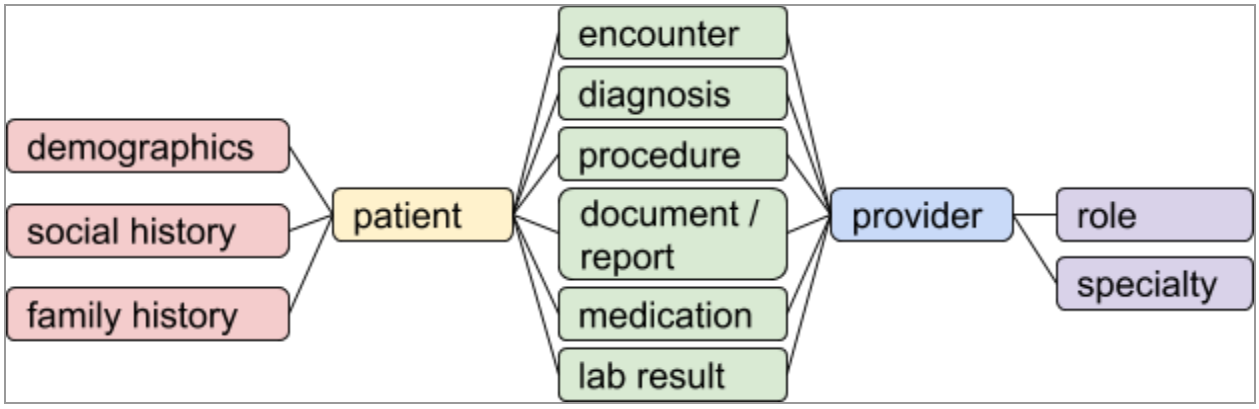

Figure 5: STARR data model

Patient information includes demographics (age, sex, race, marital status and so on), allergies, social history (e.g. smoking, alcohol and other substance use patterns), and family disease history.

Care provider information includes their clinical role (e.g. NP, Physician, etc.), their departmental affiliation and in the case of physicians, their clinical specialty.

Encounter information includes the encounter type, the patient, the care provider, the department, the date of visit, and whether the visit was considered inpatient or outpatient.

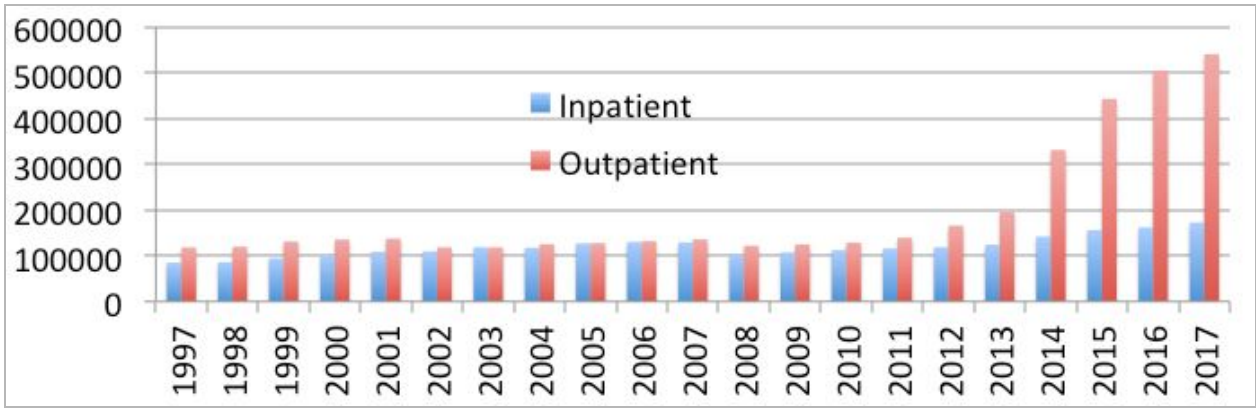

Figure 8: patient visit type count by year since 1997 from encounter data in STARR

Diagnosis codes are drawn primarily from the billing record but are also available from the clinical problem list, surgery case record, and other clinical sources. They are available both as ICD-9 and ICD-10. Over half the clinical diagnoses pertain to neoplasms, circulatory system disorders, musculoskeletal disorders and endocrine system disorders.

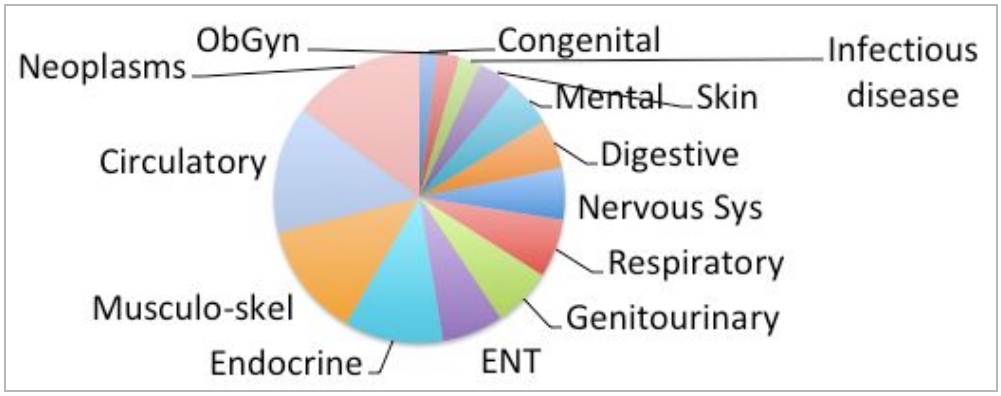

Figure 6: Diagnosis comparative frequency by category for clinical billing codes in STARR

Procedure information is primarily derived from the billing record as well, represented by CPT codes. It's interesting to see that diagnostic procedures significantly outnumber curative procedures.

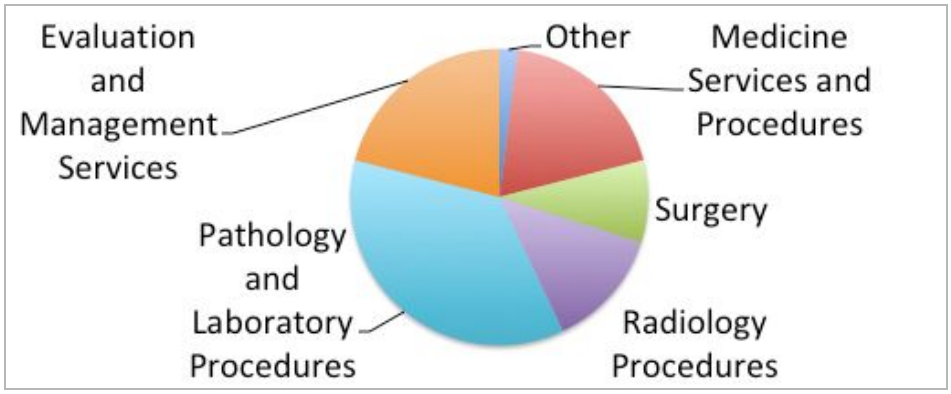

Figure 7: Procedure comparative frequency by category for clinical billing codes in STARR

Medication orders and medication administration records are documented along with the precise brand name, the dosing amount, and the frequency.

All quantitative and some qualitative Lab results are also included in this data lake.

Clinical documentation and reports remains a fascinating and rich trove of clinical information; we have included all available metadata on all clinical documents and reports, but the full text of the document or report is considered high risk data, even after anonymization. Scrubbed reports and clinical documents are available upon request to Stanford Medicine researchers with an approved IRB research protocol.

There are a few forms of data visible in Epic (e.g. Media Tab entries) that are not in the current data transfer from Hospital to School, but we can facilitate procurement of these datasets if there is a research use case. Please contact Research IT (http://med.stanford.edu/researchit/) for more information. Custom research data extracts will incur a cost, so contact us to get a cost estimate. Many research questions can be answered from the comprehensive anonymized extract known as the “STARR Tahoe” dataset group. The “STARR Tahoe” data lake contains de-identified data on 1.5M patients seen at Stanford Hospitals and Clinics and Stanford Children’s Hospital since Jan 1st 2000 and is published annually on the Population Health Sciences data portal at https://redivis.com/StanfordPHS/STARR-Tahoe/.