THE STANFORD INFORMATICS CONSULT SERVICE HANDBOOK

A guide to provide informatics consults as a clinical and research service

Contributing authors: Saurabh Gombar[1], Alison Callahan[2], Vladimir Polony[2], Keith Morse[2], Ken Jung[2], Robert Harrington[3], Nigam Shah[2]

[1] Department of Pathology [2] Center for Biomedical Informatics Research [3] Department of Medicine, Stanford University School of Medicine

1. Executive Summary

This handbook serves as a guide to creating an informatics consult service (ICS) at an academic medical center (AMC). The guide lays out the data, IT infrastructure, software, and personnel required to create such a service. In addition, there is a table of associated costs provided that can be adjusted to predict the capital expenditure and operating costs required to bring an ICS to your institution.

To begin it is important to clearly define what an ICS is, who needs an ICS and why, and expectations for successful ICS.

What is an ICS?

The informatics consult service is a way to rapidly and accurately analyze electronic medical record (EMR) data and produce actionable evidence for medical providers, researchers, and operational staff at an academic medical center. The ICS streamlines the medical data analysis pipeline from a timeline of months to days fundamentally changing the informed decision making process for those who rely on EMR data. Implemented correctly and integrated into AMC functions an ICS can improve patient care, increase research output, and reduce wasteful spending.

Need case for an ICS

Most industries have been fundamentally transformed by the ability to rapidly analyze data in order to improve decision making. Firms reinvented with on demand data analysis have created higher quality products, improved customer satisfaction, and trimmed waste all while increasing profits. Patient care, the core service of healthcare, prides itself on evidence-based decision making. By utilizing the best available evidence in caring for a patient, providers are able to reduce disease burden and improve lives. The nature of how evidence is collected and analyzed in clinical care has changed little since the analogue age.

The gold standard for clinical evidence remains the randomized control trial from which less than 15 percent of all medical decisions are based. Lesser forms of prospective evidence are derived from case-control experiments and quasi-experimental approaches; both of limited availability. The majority of clinical decisions are based on retrospective analysis of patient outcomes that are published in medical literature as retrospective studies, case series, or case reports. Despite thousands of published articles in journals every year, physicians frequently find themselves in situations where they must take an action for a patient but available evidence does not encompass the complexity of the clinical situation. Research and experience have shown that in these cases practice-based evidence, mined from the EMR, could direct clinical care. However until recently, it has been impossible to ascertain knowledge from the EMR quick

enough to be helpful in real world scenarios. The ICS generates practice-based evidence rapidly so physicians can make informed decisions for their patients.

In addition to the physicians several other groups within an AMC require large scale analysis of the EMR to perform their functions. Chief among these are those involved in medical research, hospital quality, and operations. Researchers require analysis of patient outcomes to direct research inquiries, generate hypotheses, and complete grant applications. Hard to analyze patient records results in unpursued research opportunities and fewer successful grant applications. Hospital quality and operations personnel require EMR data to track performance and optimize decisions. The ICS can generate evidence for these two groups as well to increase the success rate of their endeavors.

In all the use cases of the ICS an individual or a team is trying to answer the question “what happens to real world patients?” by leveraging EMR data. The challenges that exist currently in the journey from the question to an answer highlight why there is need for an ICS.

The first step in the journey is to acquire large scale patient data. At most large AMCs there are cumbersome ways to get patient data. These methods are cumbersome due to HIPAA compliance and large infrastructure and personnel costs required to vendor sensitive data. Furthermore, AMCs require a lengthy IRB approval process to obtain data access despite the fact that the vast majority of inquiries can be completed with a de-identified, HIPAA exempt, dataset.

Even with access to the data how does an answer seeker build a cohort of patients relevant to their question? Medical records are notoriously complex and the data required to define a cohort could exist in a medication order, textual note, disease status, procedure performed, etc. Working with multimodal healthcare data requires input from an EMR data specialist or you risk creating a biased cohort. EMR data specialists are hard to come by and have many demands on their time. Getting one for a small research or quality project is out of the question, and as a result many fruitful lines of inquiry fail.

When datasets are built and unbiased an answer seeker can move onto data analysis. Analyzing large multimodal datasets requires advanced training in biostatistics or machine learning which again are not available outside of well-funded projects. If those untrained in data science incorrectly perform analysis they frequently come up with incorrect answers leading to patient harm, institutional waste, or paper retractions.

Finally, if a question asker was able to complete this journey and put the people with relevant skills into place it can often take over a year to perform and costs thousands of dollars. Our analysis indicates the average retrospective study takes 18 months to complete. Such a timeline prohibits the use of observational data in the clinical setting and also makes it difficult for operations and quality teams to rapidly improve the institution.

In summary the current paradigm of utilizing EMR data to answer questions is slow, costly, and limited by access to experts with the skillsets and experience to answer questions from EMR data. If the challenges that are currently in place could be overcome a new paradigm for clinical care and operations emerges. The new paradigm results in precision medical care, agile institutions, and more profitable AMCs.

What does a successful ICS for clinical care look like?

For an example of the clinical value of an ICS take the case of Rosa, a 16-year old who was found to have an incidentally discovered non-symptomatic structural heart defect. She was referred to a pediatric cardiologist who had to decide, “Do I need to provide anticoagulation to this patient to reduce the risk of stroke if she develops a blood clot?” The doctor searched the literature and found only information on similar heart defects when they were severe enough to be symptomatic in infants. No published evidence was available on how to treat patients like Rosa. The cardiologist is aware that extrapolating from severely ill infants to a well appearing adolescent is less than ideal. In addition the cardiologist is aware the decision to give lifelong anticoagulation or not has very serious consequences.

If the cardiologist provides lifelong coagulation and it is not required he opens up the patient to significantly increased bleeding risk and an unnecessary cost to the payers. Similarly, if the cardiologist chooses not to provide appropriate anticoagulation where it is warranted the patient is at an increased risk of stroke leading to lifelong disability or death. How does the physician proceed? If the cardiologist could quickly determine the outcomes of all patients ever seen at his institution who presented like Rosa the cardiologist could make a better decision.

To rapidly generate this evidence the cardiologist invokes the ICS, describes the clinical context and their question to the ICS physician. The ICS analyzes the hospital’s EMR data as well as an insurance claims database of 115 million Americans to find those who match Rosa’s presentation and see if outcomes differed between patients who were anticoagulated and those who were not. The analysis takes 3 days, the same amount of time as a send out laboratory test, and the ICS provides a report to the cardiologist describing the results. Furthermore, an ICS physician discusses the findings with the cardiologist to clarify any follow-up questions. The cardiologist learns that there is no increased risk of stroke and thus decides against needless treatment for Rosa. By generating such practice-based evidence to guide his decision cardiologist was on the cutting edge of precision medicine. The decision results in improved care and reduced healthcare costs.

What does a successful ICS for quality/operations look like?

For an example of ICS utility in hospital quality consider the example of a hospitalist group involved in a cost reduction project. At their institution patients with hepatic encephalopathy, a reversible neuropsychiatric condition caused by liver failure, receive two drugs, lactulose and rifaximin. Two drugs are given not because of improved outcomes, but because published

literature demonstrates two drugs reduces the length of stay in these patients. A reduced length of stay results in higher patient turnover and a more profitable calculus for the hospital. The hospitalist spends over $200,000 a year on rifaximin, the more expensive drug, to increase overall profitability. However, the published literature demonstrating the shorter length of stay was performed in England, not at an AMC, and only on 326 patients. Thus the quality team wants to know does the two drug regimen reduce length of stay in its own hepatic encephalopathy patients. To rapidly answer this they contact the ICS which finds all patients admitted with hepatic encephalopathy who received either lactulose or lactulose + rifaximin and analyzes their length of stay. The ICS controls for all the variables leading to patient variability such as age, sex, liver function, comorbid conditions, etc. The ICS generates a report that demonstrates there has not been a reduced length of stay for patients receiving two drugs. The hospitalist group ends their practice of providing two drugs and saves the hospital $200,000 a year.

How is an ICS able to rapidly generate insight from the EMR?

The ICS leverages a multidisciplinary team and novel technology custom built to analyze EMR data.

From the need case above it is clear one of the current barriers to rapid generation of insight from EMR data is the need for specialist with diverse skills working together. The core personnel of the ICS are an informatics physician, EMR data specialist, and a data scientist.

The informatics physician has trained as a doctor, significant experience in medical research, and experience in formulating questions to be answered by the EMR. The informatics physician is ideally suited to serve as the outward facing role of the ICS. They intake all requisitions and help focus the consult requesters question. In addition, the informatics physician writes the interpretation of all reports framing them in the appropriate context.

The next member of the ICS team is the EMR data specialist. This individual has a doctorate in an informatics related field as well as years of experience working with EMR data. The EMR data specialist is an expert in building cohorts from EMR data using the custom built Advanced Cohort Engine (ACE) for search (described in section 2C). The EMR data specialist extracts the cohorts of patients data for downstream analysis. The EMR data specialist uses their experience to reduce bias in cohort creation and assure reliable results.

The final core member of the ICS team is the data scientist. This member has a doctorate in biostatistics or biomedical informatics with an extensive history publishing large scale data analysis work. The data scientist analyzes datasets provided by the EMR data specialist and generates statistics and figures. The data scientist is able to control for all the variables that bias observational studies increasing the reliability of the consult results.

The multidisciplinary team members each have complementary strengths that allow them to complete a data analysis project quickly and accurately. Furthermore, by having only three key team members interact with the patient data the ICS secures valuable patient data from security breaches.

The other key element of the ICS is the ACE search engine used to rapidly generate patient cohorts for analysis. This search engine pre-digests clinical data stored in EMR databases allowing for rapid and intuitive analysis. It is custom built for EMR data and able to handle temporal relationships between events in a patient's record. The sequential nature of patient events is why generating cohorts with conventional database technology is slow and fraught with errors. ACE speeds up cohort creation a hundredfold over standard technology.

What are the costs associated with creating and maintaining an ICS at an AMS

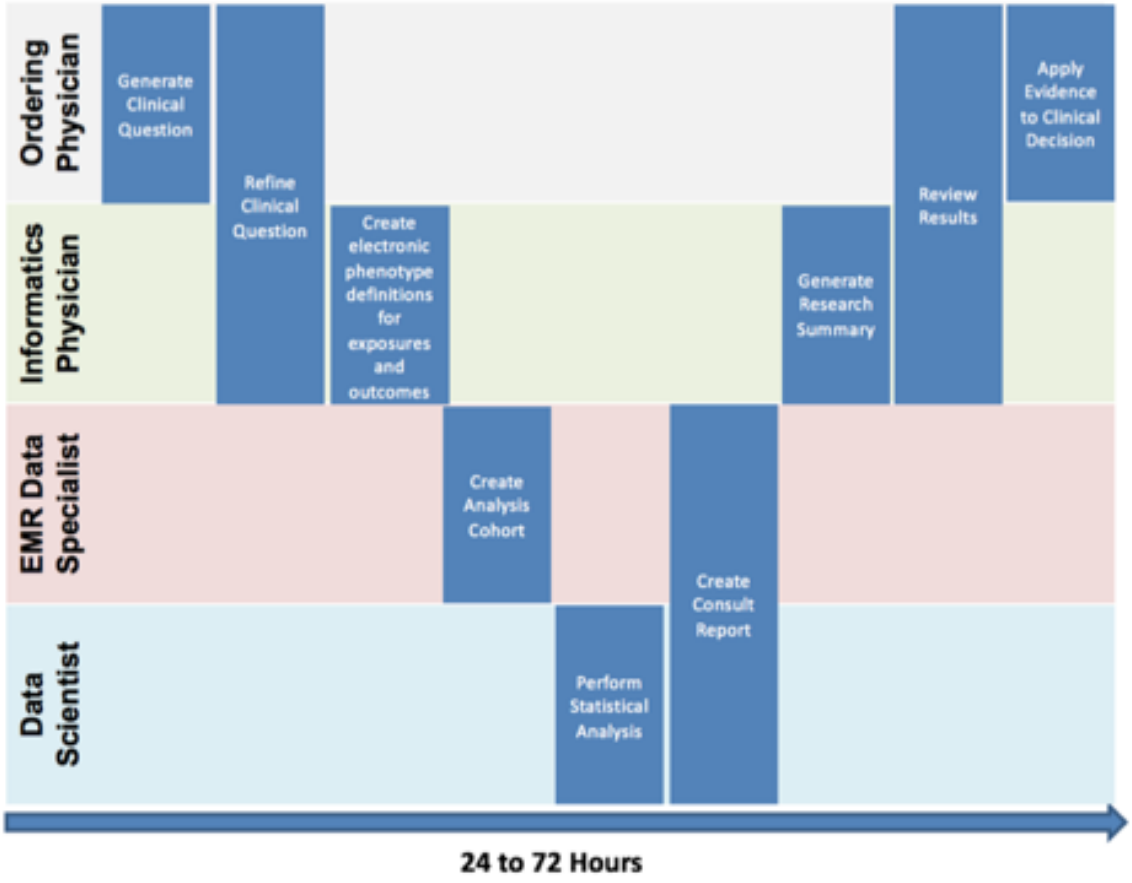
Given the large percentage of costs dependent on labor costs, it is necessary to use the pricing model in section 3.2 to accurately price the service based on the geographic area. Using costs for an AMC in the San Francisco Bay Area we project the upfront capital expenditure to be $266,00 and the annual operating costs to be $603,695 to support up to 15 queries a week. The return on investment likewise is dependent on how the ICS will be most utilized at your institution. If only answering operations and quality questions the ROI could be millions a year. On the reverse side if the ICS is answering only clinical questions it would be a net cost to the AMC. These calculations must be made in conjunction with the intended use and scope of the ICS at your institution.

2. Core ICS Components

Service Logistics

Clinicians frequently consult others in the healthcare setting with complimenting skill sets to optimize patient care. For example, a general medicine doctor will consult cardiology to help diagnose and manage a new onset cardiac arrhythmia. Clinicians have expectations as to turn around time and interpretability when they consult other specialties. The ICS , to be integrated into service must meet these operational criteria that are more stringent than a standard EMR data vendoring service available at most AMCs.

Consult workflow – The work of the consult can be performed asynchronously by the 3 core team members working on each section and passing it onto the next team member. The steps in the process are detailed in the figure below.



Interpretability – For the vast majority of cases it is not enough to simply report a numeric answer, the ICS results must be interpreted to fit the context in which the question was asked. As part of this interpretation, there should be a description of key decisions made in the analysis and any aspect of the results which might limit the finding. For example, if a physician requests "For patients with type-II diabetes who present with a hyperglycemic coma and hyperthermia does IV hydration reduce temperature as effectively as IV hydration + acetaminophen?" it is not enough to simply report, "No difference in temperature was noted between cohorts." It is necessary to describe how many cases were identified, what the average temperature and blood glucose were for the cohorts, a breakdown of demographics, the date range of mined the data, and if any unrequested but relevant key clinical endpoints differed. Without these additional points, it would be impossible for the physician to know if the answer they received is robust to their patient's situation.

Interpretations should be structured after research abstracts describing the following sections: problem, methods, a summary of results, and bottom line. Furthermore, where possible the results should be presented graphically in terms of cohort comparisons and survival analyses.

Turnaround time – To be useful for decision making the consult must have a turnaround time that is significantly shorter than the standard EMR based inquiry. The ICS should clearly state its turnaround time and like a clinical laboratory guarantee that 90% of cases will be turned around within that limit. In our experience, 3 days from intake to report could be achieved for the vast majority of questions.

Quality assurance – Given the novelty of the ICS service to medical care and limitations to observational studies, there is still moderate skepticism about use of rapidly generated patient-based evidence to aid in decision making. It is imperative that the ICS reports highlight controls that are in place in the analysis pipeline to address quality assurance. The presence of negative and positive controls (see section 2e) helps build confidence in the analyses of the primary endpoints.

Documentation/Disclaimer – When faced with a challenging scenario it is imperative that physicians document their knowledge seeking and decision making. If a physician is invoking the ICS to aid in decision making it is imperative a version controlled document be provided to them and kept on record with the ICS team. If the need exists these documents could also be scanned and placed into the patient chart to be referred to by subsequent physicians seeing this patient.

As part of the documentation, it is important that the ICS report contains a disclaimer about observational research and how such patient-generated evidence is supplemental to the due diligence of medical providers. Results in the ICS that contradict published evidence should not override established care guidelines. The suggested disclaimer for reports reads, "The informatics consult provides retrospective data analysis to support clinical decisions. The

content is not intended nor recommended as the sole basis to guide decisions. Consult published an expert opinion before committing to a patient care decision."

Secondary opinion/decision support – When physicians invoke a consult from a colleague they are looking for more than just a simple answer. Consults usually have a component of providing a platform to discuss patient details that set the scenario apart from the textbook presentation. These presentation variations are usually nuanced and best described in very technical medical terms. A discussion between physician and those not trained in medicine requires physicians to use terms understandable by a general audience often leading to a loss in nuance. As such it is important for all consult to begin with a physician/physician intake that allows the entire nuance of the medical scenario to be described. In our pilot, we found that physicians really appreciated having their question explored by another physician who was well versed the capabilities of EMR research.

Personnel requirements

The ability to deliver rapid high-quality insights from EMR data relies heavily on a knowledgeable, skillful, and well-trained staff. Complimenting skill sets amongst the team members directly corresponds to the key challenges in performing large-scale retrospective studies. These aspects include asking the correct question, extracting the correct data, performing the correct analysis, and correctly interpreting the results in the given context. The team is comprised of 3 core service providers and a single support staff. The qualifications are listed below with training details described in section 2F.

Informatics Clinician

Responsibilities: The informatics clinician will serve as the outward facing role in the ICS. They will communicate with consult requesters at intake and debrief. In addition, they will help the consult requesters refine their questions to be specific and in a format that can be answered by available EMR data. The informatics clinician is required to convert medical terms to their encoded nomenclature and write consult interpretations.

Required Qualifications:

- An MD with postgraduate training in a medical subspecialty who is still clinically active.
- A current medical license
- Ideally board certified or board eligible in clinical informatics.
- A Ph.D. or extensive experience in authoring and reviewing peer-reviewed medical literature.
- Knowledge of key medical nomenclatures including international classification of disease (ICD), common procedural terminology (CPT), logical observation identifiers names and codes (LOINC), and anatomical therapeutic chemical classification system (ATC).

- Ability to effectively communicate verbally and in writing with audiences ranging from physicians, researchers, and hospital quality/operations personnel.
- Understanding of large dataset analysis and computer programming skills.

Position Description:

Clinician with an interest in informatics and/or data science, experience treating patients, experience in EHR data analysis, and demonstrated excellence in oral and written communication (e.g. speaking engagements and peer-reviewed publications). In-depth familiarity with medical terminologies including International Classification of Diseases (ICD), Current Procedural Terminology (CPT), and Anatomic Therapeutic Classification (ATC). Experience using Epic products, including Hyperspace, and ideally Caboodle and/or Clarity. Preferable: prior experience in a data science/informatics/technology central role; proficient with R and/or Python; proficient in SQL or an equivalent database query language.

EMR Data Specialist

Responsibilities: The EMR data specialist will serve to extract relevant information from the EMR and perform all data manipulation steps. They will be responsible to generate cohorts for submitted consults, generate descriptive summaries/figures for the data, and perform basic statistical analyses on the data. They will communicate extensively with both the informatics clinician and the data scientist. This role has the largest time commitment per consult.

Required Qualifications:

- A Ph.D. in information science, data science, computer science, or related field.
- Extensive experience in authoring and reviewing peer-reviewed medical literature.
- Expert skills in the retrieval of data from the ACE search engine and relational databases.
- Ability to effectively communicate verbally and in writing
- Advanced understanding of analyzing large data sets and extensive computer programming skills.

Position Description:

Information professional/data scientist with 3+ years experience in a data science/informatics-central role. 2+ years experience conducting research using observational health data. Demonstrated excellence in oral and written communication (e.g. speaking engagements and peer-reviewed publications). Experience working in an inter-disciplinary team with medical experts, software engineers, and data scientists. Experience using medical terminologies including International Classification of Diseases (ICD), Current Procedural Terminology (CPT), and Anatomic Therapeutic Classification (ATC) to construct patient cohorts for analysis. Proficient in SQL or equivalent database query language, R (including dplyr, tidyr), and Python for data cleaning and statistical analysis.

Data Scientist

Responsibilities: Data scientist with 3+ years experience performing statistical analysis of observational data for estimation of treatment effects. 3+ years of statistical programming experience in R, and 2+ years of general programming experience in Python and using UNIX command line utilities. Ability to effectively communicate technical material to both the EMR Data Specialist and Informatics Clinician. The data scientist will be responsible for statistical analysis of the generated cohorts and developing reports of the results. They will match cohorts to control for bias and generate all statistical analyses and figures. In addition, they will develop automated tools to speed up the data processing pipeline.

Required qualifications:

- Ph.D. in data science, biostatistics, computer science, or a related field
- Expertise in data science and biostatistics
- Expertise in statistical programming in R, scripting in Python
- Ability to effectively communicate verbally and in writing

Software Engineer

Software engineer with 3+ years experience developing efficient and scalable data analysis tools in an industry setting. Expert-level proficiency with Java. Able to tackle complex problems, apply best architectural and design practices, and rapidly generate production quality code and documentation. 2+ years experience in methods spanning the entire software lifecycle, from product conceptualization, to specifications capture, functional/architectural specification, project planning, implementation, through to quality assurance. Preferable: experience developing search engine technology; experience with R/Python, including packages for data cleaning, statistical analysis, and machine learning (e.g. dplyr, tidyr, caret in R; numpy, pandas, scikitlearn in Python).

Data Requirements

Extracting, transforming, and loading EMR data for use in the ICS

The most essential piece of the ICS is access to institutional EMR data. Working with EMR data is notoriously challenging due to lack of EMR standards, a multitude of proprietary software packages, and site to site customization. The ICS is EMR provider agnostic, however, in order to take advantage of the ACE search engine (described in 2C), an EMR database ETL (extract, transform, load) is required.

The ETL needs to transform institutional EMR data into the schema supported by the ACE search engine (Appendix A). This step needs to be completed by the institutions EMR data team. The decision to strip the data of patient identifier or expose patient identifiers to the

search engine should be made at this time. Furthermore, a decision should be made about which fields should be populated for patients. Data not present in the ETL will not be able included in ICS analysis. To provide a rich consult experience we recommend the following fields as a minimum be incorporated in the ETL:

- Demographics (age, sex, ethnicity)
- Diagnoses (as ICD9/ICD10 codes)
- Procedures (as CPT codes)
- Inpatient administered medications
- Outpatient prescribed medications
- Laboratory values
- Healthcare encounters (by type of visit)
- Clinical notes
- Radiology notes
- Pathology notes
- Mortality information

The ACE search engine can handle any structured datatype as long as there is a secondary key present to attach the datatype to a patient or encounter. Additional data types that are valuable for consults include:

- Emergency room chief complaint
- Primary admitting diagnosis
- Patient vitals
- Diagnostic related groups
- Nursing flowsheet data
- Procedures (as hospital-specific procedure codes)
- Care provider details
- Patient safety events
- Press Gainey Scores
- Standard Nomenclature of Medicine (SNOMED)

Given multiple institutional demands for easily accessible EMR data, we suggest a single ETL from EMR data to the observational health data sciences and informatics (OHDSI) common data model (CDM). Several downstream analysis tools are compatible with the OHDSI CDM and ACE packages are able to automatically import data structured for the OHDSI data model. This will greatly reduce the work required to maintain multiple ETLs by the AMC's data infrastructure team.

Database administration and integrity

To provide a robust ICS the EMR data will need to be periodically updated to reflect current patient information. Given the complexity and frequent changes to EMR data structures, these updates will require dedicated personnel to assure data integrity. The percent FTE of these

administrators will depend on the frequency of the periodical update. We recommend updating the ETL at least every 6 months to capture advancements in patient care.

Furthermore, due to the changing underlying nature of EMR data structures, it is essential to have a quality control step to verify data integrity. Several key metrics should be compared for newly loaded data to the same time period over the last few years. For example, comparing the number of stents procedures or comparing a number of metformin prescriptions. The checks should encompass all data types and query frequent and infrequent encodings. These differential checks will help assure that a change to the underlying data has not left the ETL blind to important data.

ACE Search Engine

Overview - The major contributing factor to the rapid analysis of EMR data in the ICS is the utilization of the ACE search engine. ACE was custom developed at Stanford University to enable rapid creation of patient cohorts from the EMR for downstream analysis. ACE uses a temporal query language ideally suited for analyzing events which are temporally related, such as an exposure and subsequent outcome. ACE converts EMR data from a SQL relational database into a collection of patient data structures. Furthermore, ACE has its own scripting language which eliminates the need to write cumbersome and brittle SQL queries. As a result, a proficient ACE user can create cohorts in a small fraction of the time it would take to create a cohort utilizing EMR data in its native format.

ACE is available for license through the Stanford school of medicine office of technology licensing office (https://otl.stanford.edu/) as technology asset S16-392

ACE Scripting Language – The ACE scripting language allows a user to quickly create, explore, and extract a cohort for downstream analysis. To become proficient in the language takes hands-on training. For a training tutorial please see the training section below.

For more details on ACE please visit http://shahlab.stanford.edu/ACE.

Analysis capabilities

Core functionality question types addressed by ICS: The ICS service will field questions from a host of parties wishing to interrogate the EMR data to derive insight. The service is designed broadly enough to quickly answer questions in a multitude of formats. However, there are several core questions the service was designed for the reflect the nature of clinical and research queries. These question formats include:

i) “In patients presenting with X how does outcome Y differ if decision Z or Z’ is made?”

a. This question is the primary question the ICS is built to answer. Questions from physicians actively taking care of patients will often be in this form. In addition, most clinically inspired research questions will also take this form. The answer to this question will include both a summary of the data as well as statistical analysis describing the extent of outcome differences and if they are statistically significant. An example of such as question is, “In patients who present with sepsis and severe anemia is mortality improved by giving oral iron compared to IV iron?”

ii) “In patients presenting with X how many will go on to develop/receive Y?”

a. Frequently consult requesters will simply request the conditional prevalence of an outcome in a specific patient population. Since most published studies only look at very specific outcomes there are millions of disease - outcome pairs not addressed in the medical literature. In addition, quality and operations requesters will ask these sorts of question to find areas to improve. An example of such a question is, “How many patients at our institution that have stents placed are appropriately discharged on dual antiplatelet therapy?”

iii) “How many of patients at our institution received Y?”

a. Although these questions are very simple and quick to answer they are crucial for grant and talk proposals. Frequently researchers are unaware if there are even available numbers to make a research question worth the effort. An example would be, “How many patients at our institution have received therapy with CAR-T cells?” or “How frequently is non-OB surgery performed on pregnant women.”

Extended functionality question types addressable by ICS: Depending on the scope of the ICS and available personnel it can be used to answer more sophisticated answers from EMR data. These types of cases tend to require customization at the data analysis step. Examples of more involved questions include:

i) “In a patient presenting with rare finding X what are the most frequent final diagnoses they received after workup?”

a. This question is an advanced use of an ICS and makes possible a novel way to approach rare findings. Instead of measuring the frequency of a particular outcome the report generated will list all the most common future diagnoses in patients presenting with finding X. These common diagnoses tables can be used by the clinical team to tailor their diagnostic workup. An example of such a question is, “In pediatric patients presenting with mononeuritis multiform what are the most common diagnoses after 6 months?”

ii) “In patients presenting with X how does outcome Y differ from the ‘general healthy population’?”

a. At first glance, this question seems nearly identical to the first item in the core functionality list. However, this question compares a patient population to a ‘general

healthy population'. Observational studies have difficulty in comparing to a healthy population because there is no definitive time of cohort entry like there is for prospective study designs. As such to carry out this kind of consult requires the ability to do propensity score matching to create an artificial "healthy" population. This is an advanced data science technique and particularly sensitive to algorithm parameters.

iii) "In patients presenting with X which dose of Y leads to the most favorable outcome of Z?"

a. Questions about dose are difficult due to the nature of drug dosage information in the EMR. Many drugs are not prescribed in set doses but instead, doses are calculated based on weight or biological function. For example, amiodarone, a drug to treat cardiac arrhythmias, is administered as 5 mg/kg for 60 minutes. However, the EMR data from the pharmacy captures this drug only by the total amiodarone administered to the patient, the nursing flow sheet has the rate, and the doctor's note has the weight calculation. As a result, dose based questions either have to be limited to those drugs with fixed dosing regimens, ie oral tablets, or significant data clean up must occur.

iv) "In patients presenting with X how does outcome Y differ if decision Z, Z',Z'',Z''', or etc. is made?".

a. At first glance, this seems like a minor extension of the core functionality question. However, when comparing multiple treatment choices the statistical challenges are significantly increased and confounding issues are multiplied. Furthermore, pairwise comparisons of all outcomes quickly becomes difficult to interpret. When possible analyses should be limited to two cohorts.

v) "Do outcomes in hospital A in our institution differ from hospital B in our institution?"

a. To answer this question datasets from multiple institutions are required. This might not be possible for many ICS services given limited data accessibility. However, in multi-hospital systems, this type of question might be frequently requested by quality personnel trying to standardize outcome measures between hospitals within the same healthcare network.

vi) "Do outcomes for physician X differ from physician Y for presentation X?"

a. This question is not particularly difficult to execute but does require that provide identification information be provided.

Question types not currently addressable by ICS: Despite tremendous functionality in the ICS there are a number of question types and analysis types the technology and personnel cannot handle in a timely manner. These include:

i) "In patients presenting with X what is the best possible treatment choice to optimize Y outcome?"

a. Questions, where the comparators are not enumerated, are not currently possible in the ICS models. These hypothesis-free questions are of great interest to many requesters but the statistical methods and controls required to carry out such an analysis outstrip the available resources of an ICS. These questions are best referred to the quantitative sciences unit of your AMC.

ii) Situations where a cohort definition or outcome relies on natural language parsing of provider inputted free text.

a. Although notes text are part of the ICS infrastructure they are preprocessed through UMLS (discussed in a previous section). The only information exposed from the notes through ACE is the term mentions. As a result complex queries that require natural language parsing are not currently possible. An example of such as question is, “In patients with microsatellite unstable urothelial carcinomas does the rate of mitosis differs from microsatellite stable urothelial carcinomas.” In this case, microsatellite status and mitosis count both require natural language parsing of the pathology note text.

iii) Situations where cohort definitions are more specific than the nomenclature available in the data source.

a. The resolution of diagnosis codes, procedure nomenclature, and other structured medical information is at times too low resolution to capture what the consult requester needs. In these scenarios, it is preferred to refer the requester to the standard IRB protocol based inquiry of patient records.

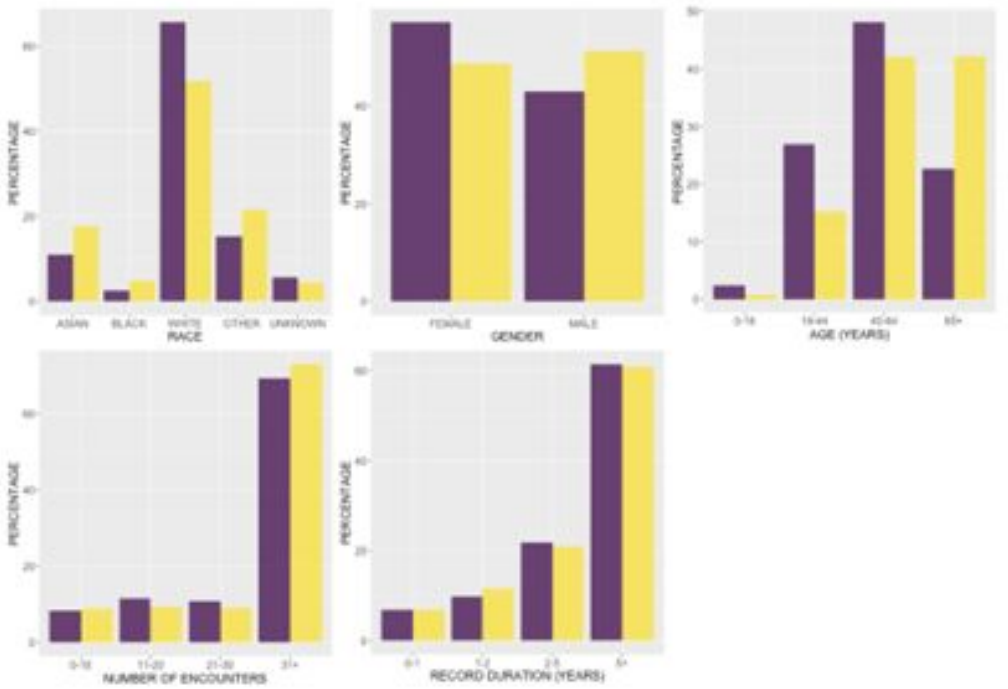
Quality Assurance

Quality assurance and reduction of bias in consults- The consult reports are effectively focused observation studies, as a result, they are susceptible to all the limitations of observational studies. Chief amongst these limitations are selection bias, missing data, and confounding. Selection bias arises from what the patient population is observed in the data set. Different hospitals have a drastically different patient population so extrapolation between hospital sets needs to be performed with caution. Missing data is a result of what information is recorded for what patients and why. Frequently patients are seen in multiple care settings or at multiple institutions. Some of the events at those institutions might not be in the dataset and this must be accounted for. Finally, confounding arises when a variable is present that influences both the cohort assignment and the outcome measured. Confounders can lead to the spurious association and causal inference methods must be chosen to alleviate confounding bias.

We recommend running a standardized methodological pipeline that controls for the largest sources of bias in the EMR. Regardless of the statistical technique used the analysis should perform on an unadjusted cohort and several different selected cohorts from the study population.

The first cohort selection should explicitly match on common confounders that frequently bias observational studies. These common confounders include patient age, patient sex, length of a medical record prior to cohort assignment event, and a number of recorded healthcare interactions. Patient age and sex confound nearly every single human disease it is essential they are accounted for when performing the observational analysis. Controlling for the number of patient encounters and length of medical record we have found does a good job to control for missing data. Patients with only one or two visits will likely be lost to follow up whereas patients with many visits are likely to be receiving longitudinal care within the system.

Figure: Example Cohort demographics:



If the consult requestor or disease pathophysiology suggests an explicit variable to match on this should be completed next. For example, a consult looking at liver failure should likely control cohorts for liver function in the form of the MELD score. Similarly, treatment response consults on diabetics should control form hemoglobin A1c values.

Finally, cohorts should be matched one-to-one on a propensity score estimate drawn from a large set of features. Such high-dimensionality propensity score systems allow for simultaneous controlling of multiple observable confounders.

Although matching can control for observable confounders, observational studies can still be biased by hidden confounders. These are confounders not measured in the dataset which cannot explicitly be controlled for. Commonly hidden confounders in EMR data include

socioeconomic status, family history, and stress. To build a level of assurance that associations drawn from consults are not affected by hidden confounders we suggest using negative control outcomes. Negative control outcomes are alternative outcomes to those of interest which are expected not to differ between the cohorts. If a negative control outcome is found to differ between cohorts it suggests a hidden confounder biasing cohort selection. For example, the administration of acetaminophen should never be causative of a temperature prior to the acetaminophen administration. Therefore to study acetaminophen's effect on temperature we can use the pre acetaminophen temperature as a negative control. A single negative control does not offer much in the way of quality assurance but a multitude of them can help strengthen the confidence in an association. The challenge then is how to pick negative controls, which usually require insight into disease pathophysiology and treatment pathways. For this, we recommend the requester and informatics clinician work together to pick several negative controls. Similarly, confounders can also cause there to be no association between outcomes where in reality there are. Positive controls, generated in the same manner as negative controls, can be used for this purpose.

For more information on optimal analysis methods to ensure high-quality results please see our publication, "Performing an Informatics Consult: Methods and Challenge [PMID: 29396125]."

Training

Consult Intakes – Consult intakes should be performed by the informatics physician as described in the personnel section. The goal of the consult intake is to capture the details of the consult question and work with the requestor to modify their question to optimize a) what they are looking to have answered and b) what can reliably answer through practice-based evidence. This process is nuanced and the best way to improve at it is to perform consult intakes. We have provided an intake script in the appendices to help direct the intake when beginning a service (Appendix 2C).

At the end of the intake, the physician should transfer the question and its details to the EMR data specialist in charge of cohort building. The transfer of this question we found was best done through a document (spreadsheet or text document) that highlights the following information:

- Population of interest
- Definition for the intervention group (or blank if an observational data consult)
- Definition for the control group (or blank if an observational data consult)
- Outcomes of interest
- Timeframe over which to do the analysis
- Data source to use for the analysis
- Variable definitions

The variable definition determination is the major time limiting step in the intake process and requires the physician find the corresponding medical nomenclature for a clinical entity. For

example, when the consult looks at outcomes for diabetics the physician must define diabetes by its corresponding ICD9 and ICD10 codes. To find the correct entity it is essential to be well versed in UMLS (https://www.nlm.nih.gov/research/umls/)

Cohort Building – The key to obtaining useful information from the EMR is to build complex cohorts correctly and efficiently. Luckily the ACE search engine with temporal search makes this step significantly easier than previously possible. However, to learn the ACE query language is a significant undertaking. We have provided an ACE tutorial in the appendix (see section appendix 2D). In addition, more details on ACE are available at https://shahlab.stanford.edu/ACE.

As with the consult intake personnel, the cohort builder must be trained in UMLS to verify the selection of codes used in the variable definitions is correct.

Furthermore, when the cohort building is complete datasets should be extracted and formatted correctly for downstream analysis. This format should be decided upon prior to beginning consults by the statistician and the EMR data specialist.

Statistical Analysis - In order to digest and the EMR data into useful snippets statistical methods must be applied to the cohorts. The statistical personnel should be familiar with advanced biostatistics and machine learning prior to beginning work on the consults. The analyses used in the consult should not be one-off analyses but all fit a prescribed pipeline which the statistician feels comfortable with.

Consult Debrief - The same physician who performed the intake should perform a debrief interview with the consult requester. This interview serves two main purposes, a) it offers a chance for the consult requester to ask any clarification questions about the analysis and b) it allows the ICS to receive feedback into where there process is working and what needs to be improved upon to meet client needs. The debrief should be tailored to the needs of the ICS. We have provided a basic debrief script in appendix section 2E.

3. Resource Requirements

Initial Capital Expenditures

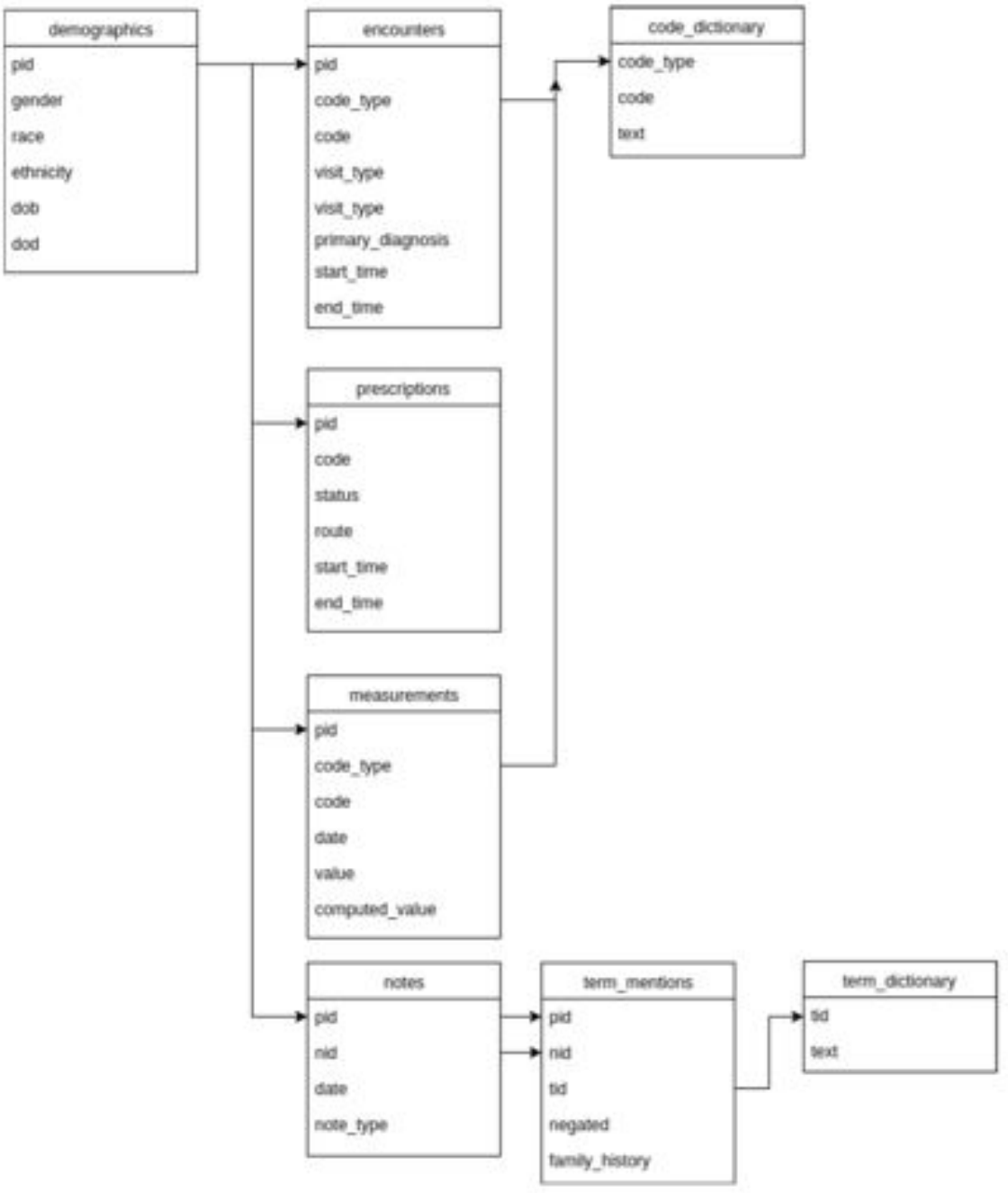
ICS Initial Capital Expenditure			
		Per employee	Total
Personnel			
	Recruitment	$6,000	$24,000
	Training/Onboarding	$1,000	$5,000
Physical Assets			
	HIPAA compliant server		$15,000
	Computers	$3,000	$12,000
Set up cost			
	Data acquisition (local)		$50,000
	Claims data		$130,000
	Server, Service, and Software setup		$30,000
Total			**$266,000**

Annual Operating Expenses

ICS Annual Operating Expenses (upto 15 consults/week)					
		Annual	Base	Benefits	FTE
Personnel		Annual	Base	Benefits	FTE
	Data Scientist - Infrastructure, Lead	$69,920	$139,840	$23,773	50%
	Software Engineer	$78,776	$157,551	$26,784	50%
	Informatics Physician	$112,500	$225,000	$38,250	50%
	EMR Data Specialist	$69,920	$139,840	$23,773	50%
	Total	**$443,695**			
Digital Infrastructure	Research IT Support	$25,000			
	Cloud Computing	$25,000			
	Data Infrastructure	$10,000			
	Professional Services	$10,000			
	Total	**$70,000**			

Physical Assets					
	Office space	**$90,000**			
Total		**$603,695**			

Appendix A: The ACE database schema



Appendix B: The ACE data model

Code dictionary:

column name	data type	required	comment
code_type	string [CPT, ICD9, ICD10]	yes	unique patient identification number
code	string	yes	case insensitive string representation of gender
text	string	yes	case insensitive string representation of race

code_type:

Demographics:

column name	data type	required	comment
pid	integer [0 - 2147483647]	yes	unique patient identification number
gender	string	yes	case insensitive string representation of gender
race	string	yes	case insensitive string representation of race
ethnicity	string	yes	case insensitive string representation of ethnicity
dob	timestamp	yes	date of birth (YYYY-MM-DD HH:MI:SS)
dod	timestamp	no	date of death (YYYY-MM-DD HH:MI:SS)

Encounters:

column name	data type	required	comment
pid	integer [0 - 2147483647]	yes	unique patient identification number
code_type	string [CPT, ICD9, ICD10]	yes	case sensitive, specifies which code type is to be recorded (CPT, ICD9 or ICD10)
code	string	yes	case insensitive code
visit_type	string	yes	case insensitive description of visit type (inpatient, outpatient, etc.)
department	string	no	case insensitive description of department
primary_diagnosis	integer [0, 1]	yes	1 = primary diagnosis (only for ICD9 or ICD10) 0 = not primary diagnosis
start_time	timestamp	yes	YYYY-MM-DD HH:MI:SS
end_time	timestamp	Yes	YYYY-MM-DD HH:MI:SS

Prescriptions:

column name	data type	required	comment
pid	integer [0 - 2147483647]	yes	unique patient identification number
code	integer [0 - 2147483647]	yes	RxNorm code
status	string	yes	case insensitive (continued / discontinued, etc.)
route	string	yes	case insensitive route (oral, iv, etc.)
start_time	timestamp	yes	YYYY-MM-DD HH:MI:SS
end_time	timestamp	yes	YYYY-MM-DD HH:MI:SS

Measurements:

column name	data type	required	comment
pid	integer [0 - 2147483647]	yes	unique patient identification number
code_type	string [VITAL, LAB]	yes	case sensitive, LAB records measurements under labs, VITAL under vitals
code	string	yes	case insensitive name of lab / vital measurement
date	timestamp	yes	YYYY-MM-DD HH:MI:SS
value	float [0 - 10000000]	no	measurement value of vitals or labs
computed_value	string	no	computed value of a lab (high, low, etc.)

Notes:

column name	data type	required	comment
pid	integer [0 - 2147483647]	yes	unique patient identification number
nid	integer [0 - 2147483647]	yes	unique note identification number
date	timestamp	yes	YYYY-MM-DD HH:MI:SS
note_type	string	yes	description of note type (progress note, etc.)

Term mentions:

column name	data type	required	comment
pid	integer [0 - 2147483647]	yes	unique patient identification number
nid	integer [0 - 2147483647]	yes	unique note identification number
tid	timestamp	yes	unique term identification number
negated	integer [0, 1]	yes	1 = term was negated, 0 = not negated
family_history	integer [0, 1]	yes	1 = term was family history, 0 = no fam. hist.

Term dictionary:

column name	data type	required	comment
tid	integer [0 - 2147483647]	yes	unique term identification number
text	string	yes	text string

Appendix C: Consult intake script

HOW TO USE: The purpose of this document is to provide a template to use during the intake stage of the informatics consult. Prior to intake make a copy of this document, add the question posed by the clinician, and remove irrelevant questions.

PRIOR TO INTAKE: Take a look at the question posed and try to determine if the key elements of PICOT are present.
Population - age, gender, ethnicity, individuals with a given disease, etc
Intervention - What identifies the individuals in your population as the study group, this could be a prognostic factor, the presence of a disease, actual medical intervention, etc
Comparison - What identifies the comparison/control group within the population. This could be all members of the population not meeting the intervention criteria or alternate stringent criteria.
Outcome - What you would like to measure between the intervention and comparison group. Could be time to death, development/progression of the disease, length of stay, etc
Timeframe - What are the time constraints on the data that are being analyzed. This includes the calendar year and even time between events to measure an outcome.
Additional Information: Any fringe cases that might arise during the analysis.

There are certain answers we cannot answer at this point. As our capabilities grow this will change but currently we have data from Stanford on the following:

- Labs (numeric lab values; normal, low, high as flagged by the machine)
- Demographics
- Encounters
- Diagnoses - ICD codes
- Procedures performed at Stanford - CPT Codes
- Vitals
- Notes (Type and Text)
- Death - obtained by the social security death records (complete prior to 2012)
- Medications and Prescriptions
- SNOMED Terms

Things we know we do NOT have:

- Chief Complaint at time of admission
- Death Data after 2012 (out of hospital)

INTRODUCTION: Thank you for using the informatics consult service. I read through the question posed in your email and would like to discuss the consult with you. This process is essential to guarantee that the data we analyze best answers the question you posed.

RESOLVING THE CLINICAL QUESTION:

From my understanding, the basic question you are trying to answer is: [PARAPHRASE THEIR QUESTION HERE AS YOU UNDERSTAND IT]. Is this correct?

[NOTES FOR ANYTHING ADDITIONAL THEY PROVIDE].

Great now I would like to go over the specifics so we can mine and analyze the correct cohorts

The population to be included in the study is [STATE POPULATION].

- Verify:
    - Population age
    - Population sex
    - Population ethnicity
    - Additional characteristics

Within the population we will be comparing the following groups:
[STATE THE INTERVENTION CRITERIA]
[STATE THE COMPARISON GROUP CRITERIA]

- Verify:
    - Intervention group and comparison group can be resolved with our EMR data
    - Temporal relationship of group criteria and population inclusion. (ie the patient should not have had the disease at their first visit)

For the study groups, we will be quantifying and analyzing [STATE THE OUTCOMES]

- Verify:
    - Temporally how outcome should be analyzed (after a given diagnosis etc)
    - The outcome is something that will be represented in the EMR data
- MAKE SURE TO ASK ABOUT CONTROL OUTCOMES

Given that we have enough patients within the data set we will perform matching between the groups to compare similar patients. By default, we will match on age and sex. Is there additional criteria you would like us to match during our analysis?
[NOTES ON ADDITIONAL MATCHING CRITERIA]

Is there anything else you think we should know to answer this question? Known confounders? Change in practice guidelines over the years 2009 and 2016?

NEXT STEPS:
That concludes the questions I have at the moment about your question. From our end the next step is coding this question into a query language and building the cohorts. If additional ambiguities arise during that step I will be in contact via email to elucidate them. Since the consulting process is new and we are still working out the process the turnaround time might be up to a week. I will be in contact in one week if we need additional time.

Appendix D: Consult Debrief script

GOAL

- Understand how information available from ACE is perceived by clinicians, and how it might influence their work.
- Identify what information is helpful for clinicians, and what information is unclear.
- Understand how to best convey outcomes and uncertainty in a clinician report.
- Get information to inform the right level of explanation for clinicians, so that they feel empowered but not overwhelmed by the information provided.

HOW TO USE THIS SCRIPT

- The debrief is intended to be performed by the same physician who performed intake
- Go to file -> make a copy
- Remove irrelevant questions (for example, if you already know details about the clinician that are being asked, or if a section was not included in your report).
- Save your copy with notes, if desired, in a separate folder.
- Color all responses in RED, please

Questions? Ask Saurabh (sgombar@stanford.edu)
Last updated: 11/3/2017

GENERAL NOTES:

INTRODUCTION
Hi, [requester's name]. Thank you for submitting your consult about [paraphrase question here]. Today I'd like to through the report and get your to feedback on each element, clarify anything you find confusing, and see if there is any follow-up questions you have. This should take 10-15 minutes and will require you to have the report open in front of you.

BACKGROUND
First, I just want to confirm some details about your background

- What's your area of specialization?
- How long have you been practicing or what stage in training are you?
- What percentage of time do you spend on research?
- How did you hear about our service?

THE CLINICAL QUESTION

- Can you share some background about what prompted that question?
- How does not having a clear question affect your patient management (skip for nonclinical questions)
- If you were able to answer your question would your colleagues be interested in that answer as a publication? Would that publication be impactful?

THE CONSULT SERVICE

Thank you for those answers. Now I would like to focus on our service and the report. Before we continue is there anything about the question or answer that I can clarify for you?

Intake

- Let's start with the conversation we had clarifying the question a few days ago. Can you give me your thoughts about that conversation?
- Do you think the conversation helped clarify and focus the question for you?
- On a scale of 1 to 5 with 5 being extremely helpful and 1 being not at all helpful, how helpful was that conversation?

How did we ask the question?

- Next, let's move to how we formulated your question to fit the data source. Do you agree with how we formulated the question?
- Is there anything you would add or remove from this section?

Researcher Interpretation

- Next, let's go to the researcher's interpretation. Can you give me your thoughts on this section?
- Do you agree with the author's interpretation of the data?
- Would you add or remove anything from this section?
- On a scale of 1 to 5 with 5 being extremely helpful and 1 being not at all helpful, how helpful is this area?

Cohort Demographics

- Next let's go to the demographics section. Can you give me your thoughts about this section?
- Is there anything you would add or remove from this section?
- On a scale of 1 to 5 with 5 being extremely helpful and 1 being not at all helpful, how helpful is this area?

Research Walkthrough

- The final section is the research walkthrough section. Can you give me your thoughts about this section?
- Is there anything you would add or remove from this section?

- On a scale of 1 to 5 with 5 being extremely helpful and 1 being not at all helpful, how helpful is this area?

SUMMARY

Now, thinking back through everything we looked at today, I'd like to ask some wrap-up questions.

- How likely would you be to use this service again?
- How likely would you be to refer a colleague to use this service?
- How would you describe this report to a colleague if they asked you what you received as an answer to your question?
- Is there anything about the report that you think a colleague might find unclear or unnecessary
- If you had to use 3 words to describe the report as you saw it today, what would those be?

NEXT STEPS

Thanks for your time. If there is anything I can clarify or if you have any follow up questions please let me know. (Might need to do a mini-intake here)