SPECIFIC AIMS

Among US-born Texans of Hispanic ancestry (7.3 million, 27% of the State's population), annual age-adjusted mortality rates kidney cancer are 1.5-fold and 1.4-fold those of non-Hispanic whites for males and females respectively [7]. Consistently with this, in my preliminary analysis (Figure 1 in Research Plan) I found an accelerated progression to metastasis among Hispanic patients at UT Health. Understanding how socioeconomic status (SES), lifestyle, interaction with the healthcare system, metabolic syndrome, and family history each contribute to this disparity will help design appropriate cancer prevention strategies, improve clinical care pathways, and target them where they will have the most impact.

I will use an inverse propensity of treatment weighted (IPTW) [8] survival model to rank the importance of several groups of possible mediators of disparity in kidney cancer progression among Hispanic patients compared to non-Hispanic whites. I will obtain the data from i2b2 [9], an open-source data warehouse used by sites in the CTSA ACT network including ours. Developed by my co-mentor Dr. Shawn Murphy at Harvard under an NIH grant, i2b2 is used to great advantage for clinical trial recruitment but wider use in population health and health services research is impeded by lack of a simple data export path for statistical analysis. I wrote a prototype app called DataFinisher [10] which bridges this gap. I will use the protected time from this KL2 to finish this DataFinisher under the guidance of Dr. Murphy (Aim-1) then extract and analyze local data (Aim-2), and then disseminate DataFinisher to Massachusetts General Hospital (MGH) where I will replicate the data collection and analysis to demonstrate the technical feasibility of using DataFinisher for multi-site data extraction and to determine the extent to which South Central Texas results generalize to a population where Hispanic patients are primarily of Caribbean rather than Mexican descent.

To guide me in my cross-disciplinary informatics software and health-services project I recruited a mentoring team of nationally recognized experts consisting of: Dr. Shawn Murphy MD (medical informatics), Dr. Amelie Ramirez DPH (disparities and cancer prevention), Dr. Ronald Rodriguez MD (surgical oncology), and Dr. Joel Michalek PhD (retrospective analysis of large data sets). They will guide me in obtaining the training and experience I will need to carry out the following Specific Aims:

- To complete my work on open source software for data-extraction from the i2b2 data warehouse.
- To use data extracted with the novel app completed in AIm-1 to test the primary hypothesis that Hispanic kidney cancer patients have an increased risk of progression to metastasis and the secondary hypothesis that a maternal history of diabetes and cancer mediates this disparity by way of metabolic syndrome.
- To deploy my software and replicate my analysis at MGH to determine applicability of the findings to a population where the majority of Hispanic patients are of Caribbean descent and in the process demonstrate feasibility of a larger multi-site study for a future grant.

At the conclusion of this study I will have a de-identified dataset representing kidney cancer patients in three health systems (UHS, UTMed, MGH) spanning the years 2013 to 2019 along with an analysis pipeline which can be used to follow this patient cohort prospectively over multiple data updates in the respective i2b2 systems. I will use the publications and collaborations developed during this KL2 to obtain funding for a continuation of this study and expansion to additional i2b2 sites in the ACT, GPC, or ARCH networks.

SIGNIFICANCE

Since 2015, the population-adjusted incidence of kidney cancer has been increasing at 1-1.3% per year[11]–[13] with 65,340 new cases expected in 2018 [14]. Mortality rates are about 1 in 3 but among US-born Texans of Hispanic ancestry (7.3 million, 27% of the State's population), annual age-adjusted mortality rates kidney cancer are 1.4 to 1.5-fold those of non-Hispanic Whites [7]. My own preliminary analysis of UT Health patient records shows that risk of progression to metastasis is worse among Latinos (Figure 1).

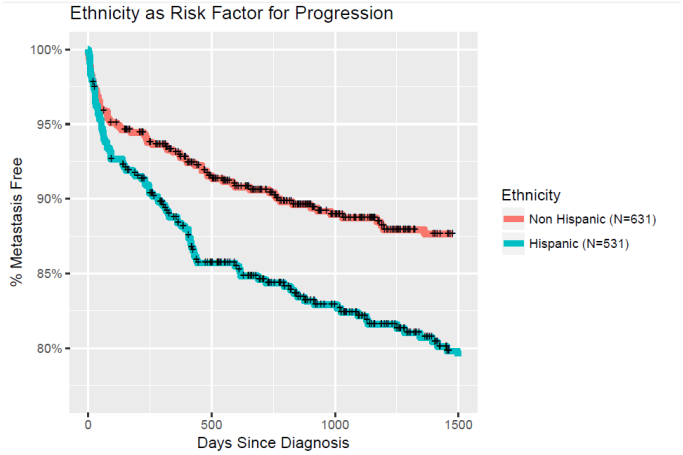

Figure 1 time until metastatic progression, Hispanic vs Non-Hispanic

There is evidence that glycolytic switch [15] is a key event in the pathogenesis of kidney cancer [16]]-- cells of the renal epithelium shift away from oxidative phosphorylation in the mitochondria toward glycolysis in the cytoplasm. This metabolic pathway is less efficient but can produce ATP and glycolytic intermediates rapidly to support rapid proliferation and resistance to chemotherapeutic agents. The glycolytic switch is triggered by increased reactive oxygen species (ROS) by way of HIF-1a. The ROS are produced by NADPH oxidases NOX4 and NOX1 in the mitochondria. These in turn are believed to be inappropriately stimulated by chronic low-grade inflammation associated with metabolic syndrome. This is where the molecular mechanisms link up to the organismal and environmental variables that I will be analyzing in order to understand what it is about about being Hispanic and living in South Central Texas that mediates the outcome disparities we and others [ref] have observed.

Dr. Rodriguez is starting a gene sequencing effort using his department's biorepository. My Aim-1 will deliver the ability to extract any needed set of variables from i2b2 for linkage to that gene expression data and my Aim-2 will suggest which variables should be linked. Biosample studies often over-represent patients of European origin [17]. My propensity scores from Aim-2 will help improve future studies by informing better sampling and weighting strategies. The major mediators identified by the model will help design and target better prevention programs and care pathways, contributing to a comprehensive model of kidney cancer that encompasses exposures at the cellular, organismal, clinical, and community scales.

EMR data can provide sample-sizes for T3 patient outcomes studies and T4 population health studies far beyond what would be feasible to accrue by any other means. The i2b2 data warehouse [9] addresses the problems of deidentification and terminology alignment and can gather records from multiple silos under a common data model. Together with a related system called Shared Health Research Information Network (SHRINE) [18], i2b2 forms the backbone of the NCATS Accrual to Clinical Trials (ACT) network. Every CTSA site that meets ACT's membership requirements has an i2b2 instance, including our own.

Use of i2b2 to get non-aggregated visit-level data for analysis is less common than for eligibility counts and other high-level descriptive results. To get the data from an i2b2 star-schema into one unified table requires a structured query language (SQL) statement with at least as many self-join clauses as there are variables. As the number of variables becomes large, such queries become complicated to write. But the obstacle to automated SQL generation has been that different variables need to be handled in different ways. One example are laboratory results with numeric values versus those with coded values (such as color, or normal/abnormal). The same variable may need different representations depending on the goals of the researcher-- some studies only need to know whether a patient is a current smoker, while others need to know whether they are a smoker at the time of the visit, or the specific type of tobacco product they consume. I am working on an i2b2 enhancement, DataFinisher, which represents a compromise between generalizability and customizability [10]. DataFinisher analyzes the properties of each variable, including cardinality and missingness and assigns the most appropriate transformation, using a list of rules that can be customized by the local informatics team. My prototype is already available to UT Health researchers with the appropriate IRB

authorization. The final result is a CSV file that can be read directly into Excel, SAS, R, and almost any analysis software but a significant amount of work remains on the rules and on preparing it for dissemination to other i2b2 sites.

By saving translational scientists the complex and error-prone chore of manually post-processing i2b2 data and offering the flexibility to amend many aspects of the dataset without re-running the query, DataFinisher can democratize T4 research and encourage wider utilization of i2b2 in health outcomes and population health studies by translational scientists who are not programmers or informaticians. This, in turn, will create additional returns on the investment that ACT sites have made in deploying i2b2.

I will be the first researcher to use DataFinisher to develop a causal model for kidney cancer based on EMR and public data sources, then test it out at a second site. This will set the stage for a subsequent study for validating this model at a larger sample of US health systems. Such a multi-site study would answer the question of whether the kidney cancer disparity we are investigating affects Hispanic persons in general or specifically those of Mexican descent-- but even the current two-site design can yield preliminary data for this question.

Clinical trials are the focus of ACT, and these too would be enhanced by streamlined data extraction. Alongside patient demographics and contact information, researchers could, for example, be provided with each patient's most recent labs, vitals, principal diagnoses, or any other supported data elements they need. Researchers could reference these columns as a cross-check during eligibility screening or use them as covariates later, when analyzing their results.

INNOVATION

For most cancers Hispanic persons have lower incidence and mortality, but for kidney cancer there is conflicting evidence. Analysis of data from the Texas Cancer Registry and the National Inpatient Sample [Michalek et al. in preparation] fails to produce evidence of disparity, while Pineheiro et al. [7] and my own preliminary analysis of UT Medicine patients suggest worse outcomes for Hispanic patients (Figure 1). Because of the size of the UHS catchment area and the fact that it is a safety-net hospital, it is a more representative sample of the Bexar County population than UT Medicine alone and can provide substantive evidence for or against Pinheiro et al. [7] at least at the regional level. Aim-3 then goes beyond the regional level, replicating the analysis at at MGH whose Hispanic population is primarily of Caribbean rather than Mexican descent. If no disparity is found, there will still be useful outcomes from the MGH arm of this study and I discuss them in the Pitfalls and Alternative Approaches section, below.

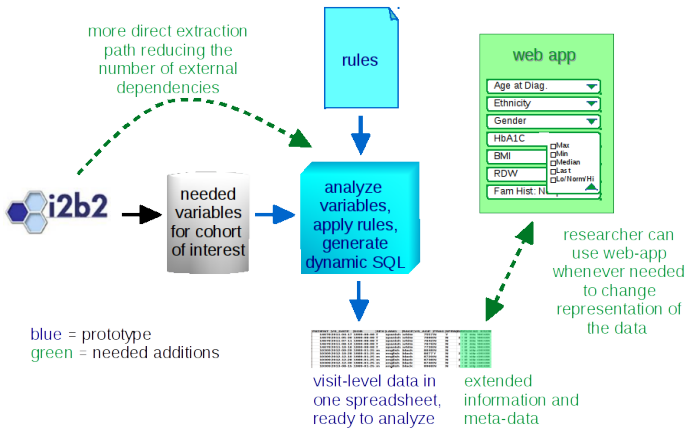

Figure 2 existing and planned features for DataFinisher

The informatics aspect of this study is innovative in that it could overturn the prevailing expectation that collection and preparation of secondary data must be a slow, labor-intensive processes and that the needs of T4 retrospective studies vary too much for automation to be feasible. My statistical training and experience with past collaborations lead me to the opposite conclusion. There is an infinite diversity of possible research plans but underlying them is a more constrained set of analysis strategies funneling into an even smaller number of expected tabular structures. Most of these, in turn, are various special cases of just one general schema similar to the "tidy data" paradigm of Hadley Wickham [19] but simpler because it only needs to encompass studies using EMR data. Briefly, each patient is represented by several rows of data, one for each follow-up period (visit, day, week, month, etc depending on the study); each data element has its own column; and patient-level values which do not change (e.g. demographics) are repeated for each row belonging to that patient. This covers survival analysis, mixed-effect models for longitudinal data, generalized linear models, linear regression, and can be easily aggregated to produce contingency tables. Embodying this approach as a tangible piece of software will help overcome the perception by many clinician scientists that EMR data is

impractical to use beyond cohort selection and feasibility counts.

I will not be the first to use causal analysis and IPTW to identify mediators of health disparities [20] but I will be the first to apply this powerful method to kidney cancer. I am proud to be an early adopter of study pre-registration, an emerging best practice to promote transparency and combat publication bias. Prior to commencing data collection and analysis I will register this study with the Center for Open Science [21].

APPROACH

Aim 1: To complete my work on open source software for data-extraction from the i2b2 data warehouse.

Initial data source on which DataFinisher will operate: The i2b2 data warehouse from which DataFinisher will pull data has de-identified electronic health records from over 379,000 adult patients going back to 2007. In 2017, the CIRD repository protocol was amended to authorize making merged data from UT Health and UHS available for research use. The initial testing is now complete and the dataset goes live this June (2018). The number of unique patients will increase to 1.2 million. My inclusion criteria are adult patients diagnosed with kidney cancer (ICD10 C64 or ICD9 189.0) followed by nephrectomy (i.e. non-surgical cases are excluded). Table 1 shows eligible patients used for my preliminary data, (i.e. not yet including UHS). For the upcoming study I will update this query to specifically select patients with renal clear cell carcinoma (RCC, which is 75% of cancer cases) but if a sufficient sample sizes are available, I may do additional analysis on papillary (10%) carcinoma. Also, I will exclude the tiny numbers of patients with HIV or with metastatic tumors *prior* to their first RCC diagnosis. The former, to avoid skewing the CCI and the latter, to insure that the index cancer is in fact the primary tumor as well as to avoid skewing the CCI (in which metastatic cancer and AIDS are the two most heavily weighted components).

	UT Health	Diagnosed	Metastatic
Total	379481	973	294
Gender			
Female	215,394 (56.8%)	399 (41%)	94 (32%)
Male	164,087 (43.2%)	574 (59%)	200 (68%)
Vital Status			
Deceased	8,177 (2.2%)	87 (8.9%)	68 (23.1%)
Living	371,304 (97.8%)	886 (91.1%)	226 (76.9%)
Ethnicity			
Hispanic	92,499 (24.4%)	392 (40.3%)	140 (47.6%)
Non-Hispanic	198,861 (52.4%)	536 (55.1%)	148 (50.3%)
Unknown/Other	88,121 (23.2%)	45 (4.6%)	6 (2%)
Age			
18-34	85,177 (22.4%)	23 (2.4%)	6 (2%)
35-44	55,622 (14.7%)	63 (6.5%)	16 (5.4%)
45-54	60,735 (16%)	166 (17.1%)	43 (14.6%)
55-64	71,031 (18.7%)	311 (32%)	98 (33.3%)
65-74	59,372 (15.6%)	268 (27.5%)	87 (29.6%)
75-84	30,094 (7.9%)	110 (11.3%)	36 (12.2%)
>=85	17,450 (4.6%)	32 (3.3%)	8 (2.7%)

Table 2 UTHealth cohort from pilot data

In addition to EMR fields (diagnoses, lab results, vital signs, medications, procedures, and demographics), our i2b2 contains death dates from Social Security records, and the contents of the local North American Association of Central Cancer Registries (NAACCR). At the time of preliminary analysis, NAACCR records were not linked to the EMR, but after the June 2018 data refresh it will be possible to match NAACCR data with EMR data for the same patients. Two other features scheduled to become available at this time are median household incomes and educational attainment from the 2016 American Community Survey (ACS) at the block-group level of precision. They will be imported into i2b2 using code I wrote earlier this year. The variables for my study will require each of these sources, and will be described in the methods for Aim 2 below.

Software development: DataFinisher is written in Python. The code will be maintained in a public GitHub repository in strict isolation from real data (even though it is de-identified) enforced by configuration settings and git commit hooks. No visit-level data will be published but the i2b2 queries producing the data will be contributed to PheKB [22] to facilitate re-use at other sites. DataFinisher documentation will generated from code-comments using Sphinx [23], a best practice among Python developers to insure documentation remains synchronized with code.

I have had a prototype of DataFinisher installed on our local i2b2 since 2015, updated as time permitted. I have three proximal goals for completing this app: streamline the variable-handling rules, retain extended data, and facilitate dissemination.

Currently, the variable-handling rules DataFinisher uses for choosing how to represent each variable are themselves complex and difficult to modify so my first goal is to generalize and simplify them.

My second goal is to implement a key new feature-- retaining additional information about each variable (including modifiers, units, and value flags). This information will be stored in specially encoded columns

alongside ordinary numeric or categoric data. Analysis software will still treat a file produced by DataFinisher as a table of data (with some additional columns of text that can be easily skipped). Yet, uploading the file back into a DataFinisher web-app which I will also develop will give the researcher a menu-driven interface to modify their variables whenever they wish, even specifying multiple derived columns for the same variable (e.g. one column with only maternal family history codes and another with only paternal ones).

My third goal for DataFinisher is making it easy to disseminate to other sites-- required for my long-term goal of leading and facilitating multi-site studies leveraging ACT and PCORI networks. I will refactor and document the installation process in accordance to i2b2 plugin deployment standards, with guidance from Dr. Murphy. As part of this, I will reduce the dependencies on external Python libraries and decouple DataFinisher from a GPC-developed app called DataBuilder [24] on which it currently relies for database connections. In Aim-3 I will describe field-testing DataFinisher at MGH.

Aim 2: To use data extracted with the novel app completed in AIm-1 to test the primary hypothesis that Hispanic kidney cancer patients have an increased risk of progression to metastasis and the secondary hypothesis that a maternal history of diabetes and cancer mediates this disparity by way of metabolic syndrome.

Hypothesized causal structure: In Figure 2 is an directed acyclic graph (DAG) with green arrows indicating hypothesized causal chains connected to Hispanic ethnicity and black arrows indicating independent effects. The dashed arrow is the direct correlation of Hispanic ethnicity with cancer progression not explained by other variables. The thick arrows represent the secondary hypothesis.

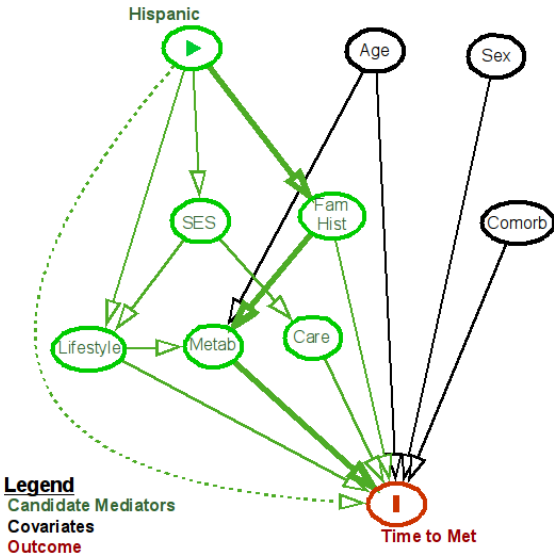

Figure 3 hypothesized causal structure

Variables: Except where indicated otherwise all variables will be from the UT Health and UHS EMR systems by way of i2b2. The main exposure will be Hispanic ethnicity. Sex and age at diagnosis will be the primary covariates. The main response variable will be time from initial diagnosis to either the first diagnosis of a secondary tumor or last tumor-free follow-up with an accompanying censoring indicator. Socioeconomic (SES) variables will include: median household income [ACS], fraction of adults completing high school [ACS], country of birth, preferred language, and type of insurance. Lifestyle variables will include smoking, alcohol, and use of anti-inflammatory drugs. Family history (Fam Hist) of neoplasia and diabetes (distinguished by maternal and paternal) will be used as a proxy for genetic predisposition. Variables related to metabolic syndrome (Metab) will include: hemoglobin A1c levels, diagnosis of diabetes, HDL, VLDL, systolic and diastolic blood pressure, and BMI. Overall comorbidity (Comorb) will be represented by the Charlson Comorbidity Index [25] calculated from problem-list codes as per Quan et al. [26]. Proxies for access to and seeking of care (Care) will be: number of visits per year, number of lab tests and imaging orders per visit, time spent with provider per visit, time from diagnosis to surgery, enrollment in adjuvant trials, and stage at presentation (NAACCR).

Statistical analysis:

Each patient in the dataset will be randomly assigned to a development, validation, or test subset. The development subset will be used to fit the propensity scores that will be described below, as an initial screen of the correlation structure predicted by the causal diagram, and for fitting the Cox proportional hazard models also described below. The validation set will be used to determine if the results obtained from the development data are repeatable. After all models and other analysis decisions are finalized, the test subset will be used to obtain significance levels and parameter estimates for publication.

To test the primary hypothesis, a Cox proportional hazard model modified to allow multiple follow-up visits per patient [27] will be fit using Sex, Age at presentation, Charlson comorbidity index, and Hispanic ethnicity as the

predictors. The response will be days elapsed from initial diagnosis to secondary tumor or last recorded visit.

If the primary hypothesis is confirmed (Hispanic variable is significant), the secondary hypothesis will be tested in stages. An ordinal variable representing maternal family history (2 = diabetes *and* kidney cancer, 1 = diabetes *or* kidney cancer, 0 = neither) will be added to the model. The hypothesis predicts that the effect of the Hispanic variable will diminish and the maternal history variable will have a significant effect. A paternal history variable will likewise be added. The hypothesis predicts that the maternal variable will remain significant. A propensity score will be constructed from the SES variables (see above) for Hispanic ethnicity. The inverse of this propensity score, the Inverse Probability of Treatment Weighting (IPTW) [8], will be used to adjust the Cox model. This will be repeated, adding variables for Lifestyle, Care, and Metabolic Syndrome to the propensity score. The hypothesis predicts that the maternal history variable will remain significant until the inclusion of Metabolic Syndrome variables in the IPTW. After that point the prediction is that the maternal history variable's effect will diminish.

If the primary hypothesis is not confirmed, the above secondary hypothesis will be modified to treat family history rather than Hispanic ethnicity as the main exposure.

Interpretation: As more variables are added, the IPTW adjusts for more and more differences between Hispanic patients and the non-Hispanic white (NHW) comparator group, so the metastasis risk attributable to Hispanic ethnicity diminishes. The fraction by which it diminishes is interpreted as the mediating effect for that group of variables. Beyond testing the causal hypothesis, the behavior of the other main effects as the IPTW term is updated will provide valuable exploratory data as a byproduct, for improved understanding of the other causal paths shown in Figure 2.

Power analysis: Time to progression will be of central importance in Specific Aim 2. Our work with existing data has suggested that time to progression (Figure 1), varies significantly with ethnicity (Hispanic $SH(4.1)=0.79$, hazard rate=0.057, Non-Hispanic $SN(4.1)=0.87$, hazard rate=0.034), where $SH(t)$ and $SN(t)$ are the progression free survival distribution functions for Hispanic and Non-Hispanic patients respectively and time is measured in years. Assuming proportional hazards, 1 year of accrual, and a total time of 5 years, and no losses to follow-up, the to attain 80% power for testing Ho: $SH(t)=SN(t)$ versus H1: $SH(t)\neq SN(t)$ with $\alpha=0.05$ the proposed study will require only N=624 patients (equal number of Hispanic and non-Hispanic patients) [PASS Version 15, NCSS Kaysville UT 2017]. The sample sizes attainable with the merged UHS/UTHealth dataset readily surpass this number.

Reproducibility and transparency: I will follow RECORD guidelines [29] in formally documenting my research strategy and will pre-register the protocol with the Open Science Framework, a practice designed to combat publication bias [21]. RMarkdown and Pandoc will be used to generate the final manuscript submissions to insure that tables, figures, and values embedded in manuscripts are automatically synchronized with the analysis results. Analysis scripts will be written using the R language [30] and versioned (separately from the data) in GitHub to insure that an audit trail exists of every modification made. The data-files will not be versioned but their MD5 hashes will be recorded whenever analysis is run, and visible in the resulting report documents.

Aim 3: To deploy my software and replicate my analysis at MGH to determine applicability of the findings to a population where the majority of Hispanic patients are of Caribbean descent and in the process demonstrate feasibility of a larger multi-site study for a future grant.

MGH: MGH, located in Boston, MA and operated by Partners Healthcare, is the teaching partner of Harvard Medical School. MGH has 1.5 million ambulatory visits and 51,000 inpatient stays per year. According to Dr. Murphy, there are a total of 18,500 kidney cancer cases in the MGH EMR. 19% of Boston's population is Hispanic. Of Hispanic Bostonians the most common ancestries are Puerto Rican (43%) and Dominican (17%), with Mexican descent comprising less than 7%.

Informatics software dissemination: If linkage to census data is not already in place in MGH, I will share with them the code we use for doing so at our site. Likewise I can share our NAACCR code if needed. I will work with Dr. Murphy and his informatics team to test and deploy DataFinisher at MGH.

Statistical analysis and interpretation: I will replicate my query at MGH, and use DataFinisher to extract a de-

identified dataset. For terms included in the ACT ontology (e.g. diagnoses) the ACT SHRINE mappings will be used to assist with replicating the query. Nevertheless, this step will likely reveal differences in how the data is coded and organized between the two sites, and I will need to do several more revisions of the DataFinisher rule file before a final dataset can be extracted. This dataset will have exactly the same structure as the one in Aim-2, but different values and a different number of results. The analysis will be as described for Aim-2 and using the same scripts but substituting in the results from MGH rather than UT Health/UHS permitting a sideby-side comparison of the sites.

Possible Pitfalls and Alternatives

Failure to confirm the primary hypothesis in Aim-2 despite a large sample size and adjustment for all relevant covariates would mean that, together Dr. Michalek's finding of no disparity in data from the National Inpatient Sample and from the Texas Cancer Registry, we would have publishable evidence contrary to Pinheiro et al. The secondary hypothesis can fail at any of the iterative steps described in Aim-2, and such failures can be informative. Briefly, part of the causal chain can be supported by the results, and the point where results are no longer as predicted is where the hypothesis needs to be updated and tested in future studies. As mentioned in the Aim-2 section, the behavior of other terms in the model can give clues about what changes are needed.

My secondary hypothesis privileges one particular set of causal relationships over many possible others. This starting model is informed by the literature about kidney cancer being a disease of energy metabolism [15], [16] and the expert opinion of my mentors about a patient's journey through the health system. I will check my results with sensitivity analysis, varying the order in which variables are added to the IPTW.

I will then perform exploratory analysis to update the causal paths for use in the confirmatory analyses of future studies. In addition what I described above, I will generate a pairwise correlation matrix of all individual data elements to eliminate non-significant relationships, shorten the list of candidate predictors, and find errors in my prior assumptions about which data elements are closely related. For example, some variables in the "Care" group are mostly influenced by provider behavior while others by the patient... if a strong clustering structure is observed, this will provide an empirical basis to separate them into two or more groups. I will also run a backwards elimination process (using Akaike information criterion) [31] to find an optimal set of main effects and interactions for predicting metastatic progression. Both exploratory analyses will be repeated at MGH and conserved features will be identified as being robust against regional differences.

Milestones/Deliverables*	**Fa 2018**	**Sp 2019**	**Fa 2019**	**Sp 2020**
Complete DataFinisher	X			
Submit Aim-1 poster to AMIA Summit	X			
Submit Aim-1 manuscript to JAMIA or JSS		X		
Pre-register research plan on Center for Open Science				
Extract and prepare San Antonio data for analysis		X		
Data analysis, San Antonio		X		
Submit Aim-2 poster to AACR, ASPC or ASCO		X		
Submit Aim-2 manuscript to Cancer Research			X	
Externship and deployment of DataFinisher at MGH			X	
Extract and prepare MGH data for analysis, do data analysis			X	
Submit Aim-3 poster to AIMA Summit			X	
Submit Aim-3 manuscript to JAMIA				X
Prepare Federal K-grant	X	X	X	X
Submit Federal K-grant			X	X

* I anticipate additional publications from this project, as well as from pre-KL2 work that I am about to submit.

Table 3 timeline for research plan and main publications. Grey 'X' represent resubmission, if necessary.

TRAINING IN THE RESPONSIBLE CONDUCT OF RESEARCH

Formal Training

I have regularly taken UT Health online training in Conflict of Interest, HIPAA Privacy Training Level 2, and General Compliance Awareness Training. I have completed CITI training in Human Research/Biomedical Research in 2015 with a refresher course in 2018. As a KL2 scholar I will continue my training by taking TSCI 5070, Responsible Conduct of Research; TSCI 6102, Practicum in IRB Procedures; and TSCI 6103, Selected Topics in Advanced Research Ethics. TSCI 5070 is a two-credit interdisciplinary course at the end of which students will be able to: (1) delineate a history of hallmark abuses of humans enrolled in clinical research, (2) describe the evolution of national and international codes and regulations guiding inclusion of human subjects in clinical investigations, (3) list the elements of informed consent and describe procedures and precautions for enrolling special populations into clinical investigation, (4) write a consent form in understandable language, (5) recognize different forms of scientific misconduct, (6) describe the role and processes of a peer review board to judge violations in research ethics, (7) develop strategies for self-assessment and validation of scientific objectivity in one's own research, and (8) recognize the ethical responsibilities and consequences of whistle blowing. TSCI 6102 is a one-credit in-depth introduction to IRB taught through a combination of readings, monthly attendance at selected IRB meetings, and discussions with faculty. I will also participate in Spotlight on Research Integrity, a monthly workshop covering current topics in responsible conduct of research. TSCI 6103 is a one-credit course where students prepare literature reviews on a topic in research ethics.

Informal Training

I will receive training in the responsible conduct of research through weekly meeting with my mentors each of which has extensive training and experience in the responsible conduct of research. The bi-weekly workshop Grant Writing for New Investigators will also provide me with training in all RCR topics as they pertain to grant application guidelines and development. I will also take programmatic RCR training offered by the KL2 program in the form of Scholars Optimizing Achievement in Research (1 hour, monthly), the Scholars Preparing Aims for R and K awards Peer-Mentoring Seminar (1.5 hours, bi-weekly), as well as monthly one-on-one meetings with the KL2 directors.

INSTITUTIONAL ENVIRONMENT

Institute for Health Promotion Research (IHPR)

IHPR investigates the causes of and solutions to the unequal impact of cancer, chronic disease and obesity among Latinos in San Antonio, South Texas and the nation. Areas of Expertise include: Latino health disparity research/training; Health promotion/communication; Cancer control research from primary prevention to survivorship; Healthy lifestyle promotion; Prevention of tobacco use, obesity, diabetes.

Masters of Science in Clinical Investigation Program

Conducted through the Graduate School of Biomedical Sciences at the UT Health Science Center, the MS-CITS degree program offers coursework and mentored research for degree-seeking and non-degree seeking students. The courses from this curriculum that I will take include: TSCI 6015 Topics in Cancer Prevention; TSCI 6065 Health Services Research; TSCI 6102 Practicum in IRB Procedures; TSCI 5070 Responsible Conduct of Research; and TSCI 6001 Introduction to Translational Science

Long School of Medicine

The Long School of Medicine, ranked one of the top three in the United States for Hispanic students by Hispanic Business magazine, has a strong and supportive faculty and numerous opportunities for building clinical and research skills. Our medical research institutes and nationally recognized cancer treatment programs combine education and research to provide some of the country's most innovative care. I will be taking the following courses here: MEDI 4153 Informatics and Advanced Evidence-Based Medicine; INTD 4104 Improving Patient Outcomes; INTD 5030 Introduction To Patient Care; ELEC 5004 Surgical Oncology Service

UHS

UHS is a nationally recognized academic medical center, network of outpatient clinics strategically located in at-risk communities, a Level I trauma center, and operates a Federally Qualified Health Center (CommuniCare). UHS is the largest Safety Net Hospital in South Texas and treats a predominately Hispanic population. Many patients are seen in UT Health clinics before and after surgical procedures at UHS.

UT Health Faculty Practice

UT Health Physicians (formerly called UT Medicine) features more than 700 physicians and health care providers offering advanced services and technologies for you and your family's needs with a patient population of over 480,000 managed via the Epic EMR system.

Clinical Informatics Research Division

The IIMS Informatics Core (B) provides EHR and billing data from our faculty practice plan (UT Health) and from University Hospital System (UHS) linked into one coherent dataset in an i2b2 data warehouse managed by the Clinical Informatics Research Division (CIRD). UT Health has been available since 2015 and UHS scheduled for production release to local researchers for June of this year, bringing the total number of unique patients from 480,000 to 1.2 million. The CIRD i2b2 data warehouse also contains mortality data from the Social Security Death Master File (SSDMF), detailed cancer reports from the North American Association of Central Cancer Registries (NAACCR) and as of June 2018 will also contain income and educational attainment data from the 2016 American Community Survey 5-year Summary linked to patients based on their addresses.

Massachusetts General Hospital

MGH: MGH, located in Boston, MA and operated by Partners Healthcare, is the teaching partner of Harvard Medical School. MGH has 1.5 million ambulatory visits and 51,000 inpatient stays per year. According to Dr. Murphy, there are a total of 18,500 kidney cancer cases in the MGH EMR. 19% of Boston's population is Hispanic. Of Hispanic Bostonians the most common ancestries are Puerto Rican (43%) and Dominican (17%), with Mexican descent comprising less than 7%.

STATEMENT OF HOW THE RESEARCH IS TRANSLATIONAL

This research falls into the T3 and T4 stages of translational research: translation to practice and translation to population health. This is translation to practice because Aims 2 and 3 involve quality improvement-- the causal network in Figure 2 of the Research Strategy funnels down into four specific and intervenable aspects of a patient's interaction with the health system at and after presentation, all of which have been selected on the basis of construct validity. Aims 1 and 3 are about dissemination of non-commercial software that will make electronic health records more accessible to researchers, as well as a demonstration project of this software.

The overall project is outcomes research that reaches all the way from the clinic into society at large where some of the underlying mediators of disparity may reside, using detailed socioeconomic data from the census linked to electronic health records.

PROTECTION OF HUMAN SUBJECTS

The proposed study merges identified data with additional sources as proxy measures for social risk factors. This research falls under the non-exempt Human Subjects Research category.

1. Risks to Human Subjects

1a. Human Subjects Involvement: Characteristics and Design: Inclusion criteria for this study are adult patients with a diagnosis of kidney cancer and who underwent nephrectomy at UT Health/UHS, or at Massachusetts General Hospital. Both sites have i2b2 data warehouses, where visit data (such as labs, diagnoses, procedures, medications, and vital signs) are stored separately from any identifying information. This enables research use of de-identified electronic health records.

There will be at least 900 subjects in Aim 1/2 of this study and at least as many additional subjects in Aim 3, the Massachusetts General Hospital arm of the study. Potential risks are limited to inadvertent release of Protected Health Information (PHI). Studies will be implemented only after final Institutional Review Board (IRB) and other appropriate regulatory committee review and approval using the SMART IRB platform.

1b. Study Procedures, Materials and Potential Risks: All analysis will be done on de-identified data. The software development activities do not require access to identified data. Actual handling of identified data will be limited to a) chart reviews to validate accuracy and b) if necessary, linkage of i2b2 patient records to sources such as US Census data or the local NAACCR cancer registry. Identifying patient data will never be transmitted between institutions and a key strategic purpose of this project is to demonstrate that a sophisticated multi-site retrospective study can be carried out without any need to transmit identifiers in the first place. Both institutions are CTSA sites with rigorous data security systems and protocols in place.

2. Adequacy of Protection Against Risks

2a. Informed Consent and Assent: This study will not actively recruit patients or use informed consent. This will be a data-only study with waived consent.

2b. Protections Against Risk: Both i2b2 teams already have established procedures in place to store identifiers on secure servers behind the firewalls of their respective with access granted only to members of the site informatics teams. Access to identifiers pursuant to this project would be limited to quality control/validation and linkage to supplementary data sources.

2c. Vulnerable Subjects: Patients in i2b2 could be prisoners, mentally disabled, have dementia or other conditions that can be classified as vulnerable but they are not being explicitly selected for these qualities. This is a data-only study; access to vulnerable subjects records will be performed with the same care as other study subjects.

3. Potential Benefits of Research to Human Subjects and Others

The researcher will not directly interact with any of the patients in the study, only patient data. Patients in the study will not directly benefit other than the knowledge generated from the proposed studies.

4. Importance of Knowledge to be Gained

While mortality rates for most cancers are declining, those for kidney cancer are increasing, and Hispanic patients are disparately impacted. This study will measure how much various candidate mediators of this disparity actually contribute to progression to metastasis among patients who were diagnosed with kidney cancer. This information can help improve the design and targeting of future prevention, screening, physician training, and health system navigation efforts. It will also help understand (in Aim 3) whether this disparity primarily affects Hispanic patients of Mexican heritage, or also Hispanic patients of Caribbean heritage. Finally, it will provide a more focused set of clinical covariates to use in future clinical trials and high-throughput analysis of biospecimens.

BIBLIOGRAPHY AND REFERENCES CITED

[1] A. F. Bokov, D. Ko, and A. Richardson, “The Effect Of Gonadectomy And Estradiol On Sensitivity To Oxidative Stress,” *Endocr. Res.*, vol. 34, no. 1–2, pp. 43–58, 2009.

[2] A. Bokov, H. Y. Liang, W. B. Qi, H. VanRemmen, W. Ward, and A. Richardson, “Diquat induced oxidative modifications track or precede cell death in mouse liver.,” *FREE Radic. Biol. Med.*, vol. 37, no. Suppl. 1, p. S105, 2004.

[3] A. F. Bokov *et al.*, “Does Reduced IGF-1R Signaling in Igf1r+/− Mice Alter Aging?,” *PLoS ONE*, vol. 6, no. 11, p. e26891, Nov. 2011.

[4] J. Bokov, Olin, Bos, Kittrell, Tirado-Ramos, “Using Prevalence Patterns to Discover Un-Mapped Flowsheet Data in an Electronic Health Record Data Warehouseigital,” presented at the Computer-Based Medical Systems, Thessaloniki, Greece, 2017.

[5] A. F. Bokov *et al.*, “Exhaustively Characterizing a Patient Cohort by Prevalence of EMR Facts: a Generalized, Vendor-Agnostic Method for Quality Control and Research,” in *AMIA Annual Symposium Proceedings*, 2017, vol. 2017.

[6] A. F. Bokov, L. S. Manuel, A. Tirado-Ramos, J. A. Gelfond, and S. D. Pletcher, “Biologically relevant simulations for validating risk models under small-sample conditions,” in *IEEE Symposium on Computers and Communication*, Heraklion, Greece, 2017, pp. 290–295.

[7] P. S. Pinheiro *et al.*, “High cancer mortality for US-born Latinos: evidence from California and Texas,” *BMC Cancer*, vol. 17, no. 1, Dec. 2017.

[8] P. C. Austin, “The performance of different propensity score methods for estimating marginal hazard ratios,” *Stat. Med.*, vol. 32, no. 16, pp. 2837–2849, Jul. 2013.

[9] S. Murphy *et al.*, “Instrumenting the health care enterprise for discovery research in the genomic era,” *Genome Res.*, vol. 19, no. 9, pp. 1675–1681, 2009.

[10] A. Bokov, L. Manuel, C. Cheng, A. Bos, and A. Tirado-Ramos, “Denormalize and Delimit: How not to Make Data Extraction for Analysis More Complex than Necessary,” *Procedia Comput. Sci.*, vol. 80, pp. 1033–1041, 2016.

[11] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2015: Cancer Statistics, 2015,” *CA. Cancer J. Clin.*, vol. 65, no. 1, pp. 5–29, Jan. 2015.

[12] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2016: Cancer Statistics, 2016,” *CA. Cancer J. Clin.*, vol. 66, no. 1, pp. 7–30, Jan. 2016.

[13] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2017,” *CA. Cancer J. Clin.*, vol. 67, no. 1, pp. 7–30, Jan. 2017.

[14] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018: Cancer Statistics, 2018,” *CA. Cancer J. Clin.*, vol. 68, no. 1, pp. 7–30, Jan. 2018.

[15] L. Yu, X. Chen, X. Sun, L. Wang, and S. Chen, “The Glycolytic Switch in Tumors: How Many Players Are Involved?,” *J. Cancer*, vol. 8, no. 17, pp. 3430–3440, 2017.

[16] K. Shanmugasundaram and K. Block, “Renal Carcinogenesis, Tumor Heterogeneity, and Reactive Oxygen Species: Tactics Evolved,” *Antioxid. Redox Signal.*, vol. 25, no. 12, pp. 685–701, Oct. 2016.

[17] A. Q. Haddad and V. Margulis, “Tumour and patient factors in renal cell carcinoma—towards personalized therapy,” *Nat. Rev. Urol.*, vol. 12, no. 5, pp. 253–262, May 2015.

[18] G. M. Weber *et al.*, “The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories,” *J. Am. Med. Inform. Assoc.*, vol. 16, no. 5, pp. 624–630, Sep. 2009.

[19] H. Wickham, “Tidy Data,” *J. Stat. Softw.*, vol. 59, no. 10, 2014.

[20] A. F. Beck, B. Huang, K. A. Auger, P. H. Ryan, C. Chen, and R. S. Kahn, “Explaining Racial Disparities in Child Asthma Readmission Using a Causal Inference Approach,” *JAMA Pediatr.*, vol. 170, no. 7, p. 695, Jul. 2016.

[21] University of Virginia & Center for Open Science and B. A. Nosek, “Opening Science,” in *Open: The Philosophy and Practices that are Revolutionizing Education and Science*, Kwantlen Polytechnic University, CA, R. S. Jhangiani, R. Biswas-Diener, and Noba Project, Eds. Ubiquity Press, 2017, pp. 89–99.

[22] J. C. Kirby *et al.*, “PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability,” *J. Am. Med. Inform. Assoc.*, vol. 23, no. 6, pp. 1046–1052, Nov. 2016.
[23] K. Rother, “Documentation,” in *Pro Python Best Practices*, Berkeley, CA: Apress, 2017, pp. 245–259.
[24] B. Adagarla *et al.*, “SEINE: Methods for Electronic Data Capture and Integrated Data Repository Synthesis with Patient Registry Use Cases,” 2015.
[25] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, “A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation,” *J. Chronic Dis.*, vol. 40, no. 5, pp. 373–383, Jan. 1987.
[26] H. Quan *et al.*, “Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data,” *Med. Care*, vol. 43, no. 11, pp. 1130–1139, Nov. 2005.
[27] P. K. Andersen and R. D. Gill, “Cox’s Regression Model for Counting Processes: A Large Sample Study,” *Ann. Stat.*, vol. 10, no. 4, pp. 1100–1120, Dec. 1982.
[28] D. A. Schoenfeld, “Sample-Size Formula for the Proportional-Hazards Regression Model,” *Biometrics*, vol. 39, no. 2, p. 499, Jun. 1983.
[29] S. G. Nicholls *et al.*, “The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines,” *PLOS ONE*, vol. 10, no. 5, p. e0125620, May 2015.
[30] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017.
[31] W. N. Venables, *Modern applied statistics with S*, 4th ed. New York: Springer, 2002.