Research plan

My best research efforts have been collaborative. I feel like these efforts represent something that neither I nor my co-authors could have developed by ourselves. I've had many wonderful research collaborations over the years. I want to highlight two of them: patient accrual in clinical trials and with mining information from the electronic health record. I also need to describe my efforts to help others be successful in their research endeavors.

Patient accrual

In 2006, I gave a journal club presentation on how to use control charts to monitor the process of patient accrual in a clinical trial. Too many studies, I hate to say, fail to meet their sample size requirements because researchers grossly underestimated the amount of time it would take to recruit patients. I discussed this control chart in the context of a clinical trial, but it really applies to any prospective research study that collects data from human volunteers.

Byron Gajewski, one of the other faculty members at the journal club, suggested that this problem might be better solved with a Bayesian approach. That turned my research around 180 degrees but it was worth it. His suggestion led to a very profitable avenue of research for the two of us.

The beauty of a Bayesian approach is that it requires the specification of a prior distribution. The prior distribution represents what you know and what you don't know about how quickly volunteers will show up on your doorstep asking to join your study. You can think of the prior distribution as a way to quantify your ignorance. You choose a very broad and variable prior distribution if you know very little about accrual patterns. This might be because you're new to the area, you're using a novel recruiting approach, and/or there's very little experience of others that you can draw upon. You choose a very narrow and precise prior distribution if you've worked on this type of study many times before, your recruiting techniques are largely unchanged, and/or there's lots of experience of others that you can draw upon. What you don't do, even if you are very unsure about accrual, is to use a flat or non-informative prior. A flat prior would be like admitting "I don't know: the study might take ten weeks or it might take ten years and I think both possibilities are equally likely." Someone with that level of ignorance would be unqualified to conduct the research.

The very act of asking someone to produce a prior distribution will force them to think about accrual, and that in and of itself is a good thing. But the advantage of specifying a prior shows during the trial itself. As the trial progresses you get actual data on the accrual rate, and you can combine that with your prior distribution, as any good Bayesian would, to get an updated estimate of how long the trial will take. Here is where the precision of the prior distribution kicks in. If you have a broad prior with high variance, then even a little bit of bad news about accrual during the trial itself will lead to a drastic revision in your estimated time to complete the trial. You will act quickly, either by adding extra centers to a multi-center trial, hiring an extra research co-ordinator to beat the bushes for volunteers, or (if the news is bad enough) cutting your losses by ending the trial early for futility. If you have a narrow prior with low variance, then you've done this trial often enough that you don't panic over a bit of bad news. If the data keeps coming in and it shows a much slower accrual rate than you expected, then you will eventually reach a point where you need to take action. But there's a cost associated with a premature overreaction that a precise prior will protect you against.

Dr. Gajewski put one of his graduate student, Joyce Jiang, on the trail, and she contributed several additional publications after completing a successful dissertation defense of her extensions in this area. I worked very closely with Drs. Gajewski and Jiang, and found an interesting theoretical contribution to Bayesian data analysis that was hidden in their work.

One of the problems with getting researchers to produce a prior distribution is that they sometimes are wrong–spectacularly wrong. If you have a strong prior attached to a prior that is sharply at odds with the

actual accrual data, you'd like to find a way to discount that prior distribution, but you'd like to keep that strong prior for the precision it gives you when the prior and the actual accrual data agreed with one another. They came up with a very clever solution. Attach a hyperprior to the precision of the prior distribution. If the accrual data and the prior are in sync, the precision stays high. But if there is a serious discrepancy between the accrual data and the prior, the hyperprior shifts and leads to a much weaker prior distribution.

I dubbed the method they proposed the hedging hyperprior, and suggested that it might work in other Bayesian settings as well. It turns out to be equivalent to the modified Power prior proposed by Yuyan Duan in 2006, but the formulation of the hedging hyperprior is both simpler and more intuitive. I have presented a simple example applying the hedging hyperprior to the beta-binomial model and am preparing a manuscript for publication.

The work that Dr. Jiang did on her dissertation was not just limited to the accrual problem but included an additional Bayesian application to a validation model using expert opinion. The strength of her work in these two areas led to her appointment to a post-doctoral fellowship at Yale University. Dr. Gajewski and I continue to collaborate with her on these models. We have four peer-reviewed publications and an R package so far, and plan to collaborate with other researchers in this area.

Closely related to my research on patient accrual is an effort to audit the records of Institutional Review Boards (IRBs). Too often, researchers fail to obtain the sample size that they promised in the original research protocol, mainly because subject recruitment takes longer than expected. It is very easy to compare the protocol submitted to the IRB to the final report. In the study of 135 submissions to one IRB, more than half of the researchers failed to reach their enrollment targets and the average shortfall was more than 50%. I have approached many other IRBs to ask to replicate this work, but none have shown any interest. But I plan to continue to ask anyone who works on an IRB to help me.

Mining the electronic health record

In January 2016, I was offered the opportunity to work on a research grant funded by the Patient Centered Outcomes Research Institute. The grant supported the Greater Plains Collaborative, a consortium of ten academic health care centers in the midwest. It was run out of Enterprise Analytics, located at Kansas University Medical Center. I jumped at the chance and dropped much of my other work to focus on this grant.

My assignment, derived through discussions with the head of Enterprise Analytics, Russ Waitman, was to develop a develop a phenotype of breast cancer from information in the electronic health record (EHR) and validate it against information in the breast cancer registry.

Such a phenotype would have great value in identifying patients for prospective clinical trials. The advances in high throughput genome sequencing and the linkage of that information with the EHR allows for exploration of novel precision medicine options. Developing a phenotype from the EHR is fraught with peril because information in the EHR on basic issues like diagnoses and treatments is often coded inconsistently. Having a link between the EHR and a tumor registry provides an external validation of the accuracy the EHR phenotype.

The EHR at KUMC (and the other sites in the Greater Health Collaborative) is stored in an Oracle database and is accessible through an i2b2 system that is very easy to use. My work, however, required access to the full database as well as some of the metadata. With the help of Dr. Waitman and the SQL expert sitting across from me, Dan Connolly, I was able to pull information directly from Oracle.

I used a big data model, LASSO regression, to predict whether a patient was in the breast cancer tumor registry and set up sparse matrices as input to better manage the size of the data sets. The breast cancer cases were compared against three separate control groups and in spite of the massive size of the independent variable matrix (more than 45,000 columns), this model ran in under ten minutes. The resulting sensitivity and specificity were very high, putting to rest concerns that the EHR data might be too incomplete or inconsistent to produce an accurate phenotype. The LASSO regression model could easily be run for other

tumor types, and just as quickly validated. I have presented these results at a local research conference and plan to submit a peer-reviewed publication soon.

An interesting side effect of this research is also worth mentioning. I had a passing knowledge of SQL prior to my work with Dr. Waitman and Mr. Connally, but I had to quickly learn and apply a broad range of SQL programming to get the data into R and the LASSO regression model. SQL is a fairly easy language to learn, but many of the students in the Bioinformatics program at UMKC do not have even a cursory knowledge of SQL. So I am partnering with a database expert at UMKC to develop a new class, Introduction to SQL, that will cover some of the basic skills that an researcher would need to query data stored in a relational database. This class is not intended to teach someone to become a database administrator, but rather a competent user of other people's databases. It will parallel to a large extent classes that I have already helped develop and teach: Introduction to R, Introduction to SAS, and Introduction to SPSS.

My work on the PCORI grant has been transferred a different grant and I plan to work with partners at Truman Health Center and Children's Mercy Hospitals to develop more research utilization of EHR at their locations. I also have been asked to help develop an analytics platform simplifies data mining of the EHR through a standardized library of functions interfacing SQL databases and R. This library would pull appropriate meta data descriptors as well, expanding the types of analyses available to the end user.

Helping others with research

I am a great believer in the Harry Truman quote "Anything is possible if you don't care who gets the credit." A major portion of my career has been helping others become successful researchers. This work is quiet, behind the scenes, and often leads to very little recognition for me. But I enjoy watching someone developed from a scared and timid person starting out with their very first research study to someone who has learned enough that now he/she is mentoring others.

A large part of my work is helping people who are struggling in their graduate studies. It might be some extra tutoring for that seemingly impossible statistics class. More often, though, it is guiding students through the difficult process of writing and defending their dissertation. I did this for free for doctors, nurses, and other health care professionals that I worked with at Children's Mercy Hospitals and Clinics. After I left that job, I started my own consulting business, and I got lots of referrals to graduate students. They typically are struggling with their disserations and with a dissertation committee that was not giving adequate direction on the data analysis. For a dissertation, you can't do the data analysis for them because they have to be able to explain their work during the dissertation defense. You have to teach them those things that they didn't pick up in their earlier statistics class and teach them so thoroughly that they can offer a clear and convincing presentation of the analysis that they ran. You have to help them anticipate the types of questions that they might get and how best to answer those questions. Perhaps the most important thing is to get them a sense of self-confidence that they are working on an important problem and that they have a solid and defensible plan for solving that problem.

Another big portion of my work is helping people write their first grant. The first grant is almost always for an amount too small to support a statistician as a co-investigator. But these are the grants that need to most help and support. I help with the research design, the sample size justification, and the data analysis plan. But my work is not just limited to that. If the specific aims are vague, if the literature review rambles incoherently, or if the research budget is unrealistic, I offer gentle suggestions to fix these problems. I try very hard not to overstep my bounds; I don't know the science as well as they do. But I can often provide valuable feedback by looking at things from the perspective of an outsider who has seen hundreds of grants in dozens of different scientific fields.

I've taken many classes in grant writing to better understand the whole process and to improve my ability to work with new researchers. But recently I have taught the classes myself. In 2012, I co-developed a class in grant writing for researchers in Complementary and Alternative Medicine with a prominent statistician working for what was then called the National Center for Complementary and Alternative Medicine and with another statistician working at a major college of Chiropractic Medicine. We repeated this class in 2014 for the same conference. I provided lectures on designing a pilot study, justifying your sample size, and what a

scientist should look for in a collaborating statisticians. I gained the most benefit, however, by listening to the lectures of the other statisticians. One of the best was a discussion about insuring consistency between the specific aims, the research design, the data analysis plan, and the budget. Another excellent talk was on hiring a data manager and preparing a solid data management plan (including budgets!).

When you include the small internal grants that provide seed money for new researchers, I have helped write hundreds of grants, too many to track. I do list those grants where I am on the budget, but these represent just the tip of the iceberg.

Finally, I help new researchers navigate the daunting process of getting Institutional Review Board (IRB) approval for their studies. At Children's Mercy Hospital and Clinics, I was housed very close to the people who ran the IRB and got to know them very well. It would have been a conflict of interest for me to serve on the IRB, because I would be reviewing protocols that I helped write. But I did work extensively with the IRB, answering technical questions, accepting referrals from researchers whose protocols had glaring problems with scientific rigor, and providing training courses on ethical issues associated with research. Statisticians, I believe, are key players in assuring the ethical conduct of research. Often researchers are barred from the optimal research design by ethical constraints and our job is to help find an alternative design that meets the needs of the researcher while still protecting the rights of the research volunteers.