

A Day in the Professional Life of a Collaborative Biostatistician Deconstructed: Implications for Curriculum Design

Gregory P. Samsa^{1,*}

¹Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

*Correspondence: 11045 Hock Plaza, Durham NC 27710, USA. Tel: 1-613-5212. E-mail: Greg.Samsa@duke.edu

Received: December 22, 2017

Accepted: January 3, 2018

Online Published: January 17, 2018

doi:10.5430/jct.v7n1p20

URL: <https://doi.org/10.5430/jct.v7n1p20>

Abstract

Collaborative biostatistics is the creative application of statistical tools to biomedical problems. The relatively modest literature about the traits of effective collaborative biostatisticians focuses on four core competencies: (a) technical and analytical; (b) substance-matter knowledge; (c) communication; and (d) problem solving and problem framing. Most statistical education concentrates on the technical and analytical competency; here, we focus on the remaining ones. Case studies describing consultations about study design and data analysis are presented, and the task is to deconstruct the knowledge used by an experienced collaborative biostatistician into components which are more explicit (and, ultimately, teachable). These components include specific and concrete information about statistical procedures; substance-matter knowledge about biology and medicine; general knowledge about biomedical studies, especially study design; insights about the process of effective collaboration; and high-level synthesis. Implications for curriculum design are discussed. To follow up on these qualitative and provisional efforts, the next step in scholarly research about to teach communication, problem framing and problem solving within the context of collaborative biostatistics should focus on a finer-grained and evidence-based description of what these competencies actually entail.

Keywords: *collaborative biostatistics; curriculum design; competencies; case study*

1. Introduction

Collaborative biostatistics is the creative application of statistical tools to biomedical problems – in other words, “applied statistics” where the area of application is biology and medicine. I’m a collaborative biostatistician and, indeed, have been practicing this art in a major research university for more than 30 years. (From now on statistics and biostatistics will be used interchangeably).

The literature on what makes an effective collaborative statistician is remarkably modest (e.g., Baskerville, 1981; Begg & Vaughan, 2011; McCulloch *et al*, 1985; Sharples *et al*, 2010; Taplin, 2003; Taplin, 2007; VanBelle, 1982; Winkvist, 1990). “What makes an effective collaborative statistician” can be conceptualized around traits (i.e., distinguishing qualities) or competencies (i.e., the ability to do something successfully or efficiently). Both are important: for example, during the admissions process we might attempt to identify candidates having the trait of flexibility of mind and also the competency of mastery of advanced calculus. Competencies, which are the focus here, are particularly useful for curriculum development and evaluation: for example, the more precisely we can define what it is we want our graduates to be able to do the more likely it will be that we can design the curriculum to achieve that goal, and the more straightforward it will be to assess the degree to which that goal has been achieved.

Even though the competencies of effective collaborative statisticians are described in the literature using varying terminology, these conceptualizations are fundamentally similar and also consistent with experience. Here, and noting that the final two competencies are sometimes mixed together, we consider four core competencies: (a) technical and analytical; (b) substance-matter knowledge; (c) communication; and (d) problem-based (i.e., problem solving and problem framing). Technical / analytical competence includes, among others, familiarity with various statistical approaches, understanding their assumptions and mathematical basis, and skills in applying statistical software. In this particular context substance-matter knowledge primarily pertains to biological and medical content. Communication includes understanding and being understood – this being critical to effective interaction between

disciplines in general and the practice of collaborative statistics in particular. The problem-related competencies especially include the ability to identify the critical features of the phenomenon being studied, since statistics simplifies reality by focusing on some features to the exclusion of others (and it works best if the focus is on the most critical ones).

Paraphrasing an investigator when asked about the characteristics of an effective collaborative statistician: “It goes without saying that my statistician knows what to do in terms of selecting the proper statistical test. She’s also taken the trouble to learn about my topic area and this helps her to identify what’s actually important. She’s able to bring in information from outside statistics – for example, from other investigators and the literature – and also communicate information about statistics in plain English. I’m confident that we understand one another and thus that she’s doing the right things.”

Figure 1 describes a simplified model of the interactions between collaborative statisticians and other investigators. It also illustrates the responsibilities and reciprocal relationships between collaborative statisticians and others.

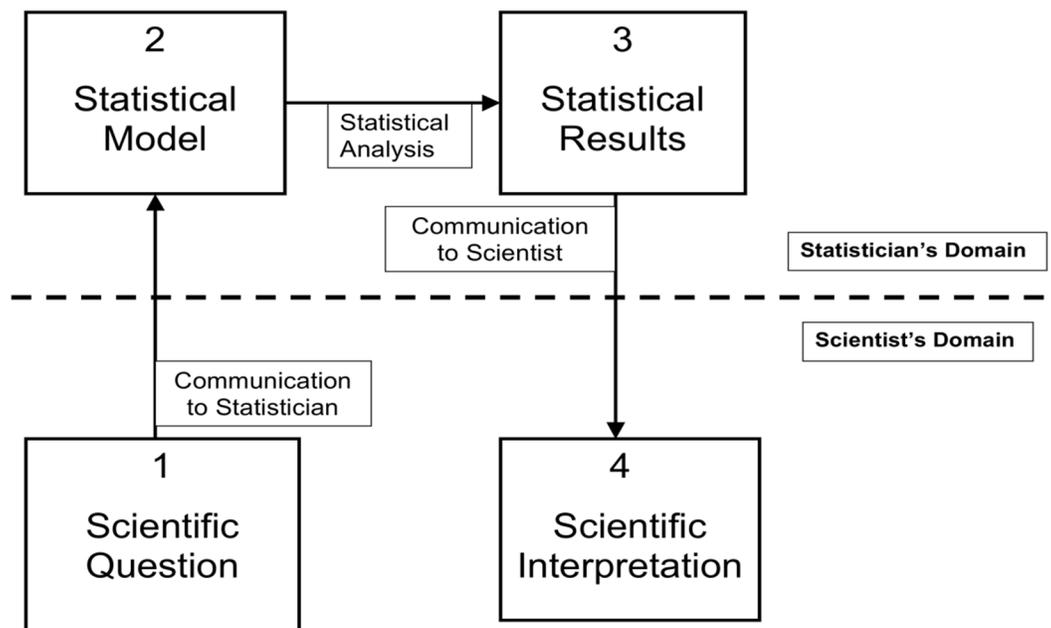


Figure 1. A Simplified Model of Collaborative Statistical Practice

Figure 1 illustrates relationships between collaborative statisticians and other investigators.

Some years ago I was tasked with developing graduate programs for my department which, although it goes by another name, is effectively a department of applied biostatistics, within which a majority of the faculty (not to mention staff) serve as collaborative statisticians. These programs don’t solely aim to train collaborative statisticians, but that’s a significant part of their rationale.

Program development began with a vision-setting exercise asking “what are the characteristics of an ideal graduate?” Sources included a review of the literature on the characteristics of effective collaborative statisticians, open-ended interviews of well-regarded collaborative statisticians and their non-statistical colleagues beginning with the question “what makes an effective collaborative statistician?”, a review of similar programs, and reflections of members of the curriculum development committee on their own experiences as graduate students.

Information from these sources was quite consistent, and focused on the above four competencies (variously defined). Indeed, one of the most notable threads of reflecting on our training as students was the disconnect between our professional experience, in which all four competencies are important, and experiencing curricula which very strongly focused on the technical and analytical competency. The other three competencies, although important, received less attention.

Our hypothesis is that one of the main reasons that these other three competencies receive less attention is that they are less straightforward to teach than the technical / analytical competency, and moreover that separating these competencies into sub-competencies can help provide a way forward. The current effort is a preliminary and qualitative attempt to identify these sub-competencies.

2. Methods

Two case studies are presented, and the task is to deconstruct the knowledge used by an experienced collaborative biostatistician into components which are more explicit (and, ultimately, teachable). These cases were selected because they represent two areas where it is generally acknowledged that experienced collaborative biostatisticians add value to the research enterprise, namely: (1) in study design, to increase the chance that a grant application will be funded; and (2) in data analysis, to increase the chance that a manuscript will be published. The case studies begin with a description of an interaction between an investigator and a collaborative biostatistician, followed by a deconstruction of that interaction.

The deconstruction of the interaction between the collaborative statistician and their colleague focuses on the “information” used by the statistician. This information could be specific and unequivocal – for example, the formula underpinning a statistical test. This information could also include general principles, rules of thumb, and the like. The former type of information is straightforward to teach, the latter type of information less so. Indeed, this latter form of information eventually becomes internalized by experienced collaborative statisticians and helps to structure how they conceptualize their discipline. These conceptualizations have some level of variation across individuals but, nevertheless, are sufficiently similar to provide common ground. Indeed, in the spirit of constructivism, part of our task as educators is not just to transmit facts, but to provide experiences that help students develop their own soundly-derived mental maps.

Finally, the elements of these interactions are tentatively grouped into sub-competencies.

3. Results

3.1 Case Study 1: Study Design

3.1.1 The Interaction

As background for a meeting an investigator has emailed a draft grant application. She has developed a new app to assist in the management of cancer-related pain. Rather than acting as a hammer in search of a nail, she plans to begin the proposed study by interviewing patients, providers and administrators to (a) fine-tune the app; and (b) determine the optimal way to embed the app in the clinical workflow (in other words, to determine how the app might be best utilized by all concerned).

The investigator proposes to test the app using a randomized trial which randomizes patients into three groups representing low-, medium- and high-intensity rollout of the app. For example, the high-intensity version of the rollout would require the greatest investment of provider time. The primary outcome will be pain control as quantified by a pain score (scaled 0-10) which is already recorded at each visit; secondary outcomes include app-specific measures such as usability. Although the investigator has pilot data about pain scores there is little to no information about how much the app is likely to improve pain control. The sales pitch for the grant is that the above level of “customer engagement” is unusual and, indeed, the combination of this with efficiently embedding a randomized trial within the clinical workflow would be at the frontier of research about how to implement “learning health care”. The question posed by the investigator was “how many patients do I need in each of the three study groups”?

The consulting session began by asking the investigator to consider some more fundamental design questions, beginning with the strengths and weaknesses of the parallel-groups randomized trial being proposed here. Pre-post designs (and variations such as a stepped wedge) were considered, most of the discussion of which centered around the question of “contamination” – that is, whether or not the clinicians would apply information from treating patients receiving one version of the app to others. Cluster randomization was also considered. Without taking this case study too far afield, it was ultimately decided that a parallel group randomized design was appropriate.

The next question considered the number of groups, with the statistician recommending two groups rather than three. The reasoning was two-fold. First, the research hasn’t advanced to the point where subtle distinctions between different versions of the app should be the primary study question – more typically, in app development the first study is usually designed as a proof of concept study that compares the best/maximal version of the app with the

weakest possible version of the app consistent with sound science. Second, as a general rule it is more efficient statistically to use two groups rather than three. The investigator stated that this recommendation would simplify a number of aspects of the study.

The statistician then suggested that the best/maximal version of the app might be based upon the interviews from providers and administrators, and be the answer to the question “what’s the most you would be willing to do (e.g., in your clinical practice)?”. The weakest version of the control group is “no app at all”. The statistician inquired about the current state of the science in this area of research as pertained to the selection of a control group, and a potential control group was identified. (For the present purposes the precise definition of the control group isn’t critical.)

The statistician suggested that the sample size be in the range of 100-200 patients per group, based upon generic effect sizes, which don’t require specific information about the expected means and standard deviations of the pain scores in the two groups, and instead summarizes information about efficacy on a relative scale. Roughly speaking, a “small” effect size of 0.2 yields a sample size of 380 per group, a “moderate” effect size of 0.5 yields a sample size of 63 per group, with 100-200 per group landing in the middle. (Cohen, 1988) These sample sizes don’t include the impact of attrition. The effect sizes for successful apps tend to range in the small to moderate range -- the best guess might well be “small”, but it’s also quite likely that in the interviews the administrators will say that they’re only willing to take action if the effect size is somewhat larger, and an effect size of 0.35 might be a reasonable compromise. The investigator noted that a study of this size would be logistically feasible without having to recruit additional sites and would be compare favorably with the studies in the literature. The next step, not discussed here, was to fine-tune the sample size calculation to arrive at a specific number.

3.1.2 Deconstructing the Interaction

The list below summarizes some of the elements of statistical knowledge which the consulting statistician applied. These elements were combined into framing, and then ultimately solving, the problems at hand.

- Although investigators often pose a specific question, the answer to that question isn’t necessarily what they need.
- Identify the study question, and also why that question is important. Study designs aren’t developed in isolation, but instead within the context of a study question.
- Begin consultations by reviewing the perceived rationale for the study design. One of the initial work products should be an explicit statement summarizing this rationale.
- The types of study designs used for similar assessments include a non-randomized pre-post design, a parallel-groups design with randomization at the level of the patient, and a cluster randomized design with randomization at the level of the cluster. Strengths and weaknesses of these designs should be considered. Ideally, knowledge of these strengths and weaknesses shouldn’t only be hypothetical, but based on experience with grant proposals that are (and are not) funded.
- In general, it is more efficient to compare two groups rather than three, especially if the middle group (whose outcomes will be somewhere in between the two extremes) is the one which is dropped.
- The natural history of research on the efficacy of interventions typically begins with a proof of concept study, which is designed to have the maximum possible efficacy, and then proceeds to studies which make subtler distinctions. Indeed, this is often what funders expect.
- Scientific reviewers pay particular attention to the justification of the control group. One option is usual care. Another option enhances usual care – for example, usual care plus an app that provides general information about pain management but isn’t tailored to the patient’s specific circumstances. Comparing against usual care maximizes the relative impact of the intervention and thus the statistical power. Comparing against a non-tailored or low-intensity version of the app strengthens the inference that what is causing the difference between groups is the specific characteristics of the high-intensity version of the app rather than the less excellent characteristics of disorganized usual care, but on the other hand the relative impact of the intervention is less, as is the statistical power.
- In general, base power calculations on the primary hypothesis only, as this streamlines the process of coming to a final decision. Secondary hypotheses either have sufficient power or they don’t. If they do, this strengthens the argument. If they don’t, these analyses can nevertheless be considered as “exploratory”.
- The technical mechanics of a power calculation require specifying (among others) the design, the primary outcome variable, the study question, the statistical test, and the inputs to that statistical test. For

continuously-scaled outcome variables such as pain scores this typically requires specifying the means and standard deviations of the study groups.

- The means and standard deviations in a standard power calculation can be translated into an effect size.
- Power calculations can be based upon the effect size that is likely to occur or the effect size corresponding to the minimum clinically important difference between the groups (here, the level of efficacy that administrators would require in order to take action). Often, the latter is preferable.
- The precedent in studying apps specifically, and implementation science more generally, is that effect sizes for successful interventions tend to range from 0.2-0.5, often toward the low end of the range.
- The precedent in studying apps is that randomized trials have 100-200 patients per group or smaller. The smaller studies tend to be negative and probably are underpowered.

3.2 Case Study 2: Data Analysis

3.2.1 The Interaction

This is an unusual study design. A large longitudinal database (20,500 records) of 6,000 patients with cancer was queried. This database contains a pruritus (i.e., itching) score (on a scale of 0-10), obtained as a matter of course during the clinical encounter, for all 20,500 observations. (Itching is a possible side effect of various cancer treatments, and if severe can have a significant impact on quality of life.) For those 500 observations (from 200 patients) with severe itching (i.e., scores of 4 and above, based on the cut-point used in various practice guidelines) additional data were abstracted from the medical chart.

The additional data collection included a list of current medications, whether any discussion of itching was recorded, and whether action regarding itching (e.g., a change in the medication regimen such as the prescription of an antihistamine) was taken. Medications were classified into two groups. In group A at least one of the medications that the patient was receiving was thought to have a significant risk of itching. In group B there were no such medications.

To help visualize the database associated with the study, the data array of itching scores for three patients with 2, 4 and 8 observations, respectively, is as illustrated in Table 1. Observations with additional data collection are indicated by an asterisk. Please see Figure 2 and Table 2 as well.

Table 1. Data Array of Itching Scores

Patient 1: 2, 1
Patient 2: 2, 4*, 5*, 3
Patient 3: 0, 0, 0, 0, 0, 0, 0, 9*

Table 1 illustrates the longitudinal nature of the data, and also that additional data collection was limited to patients with itching scores of 4-10.

Table 2 illustrates the data array in case study 2. The gender column is intended to stand in for gender and other clinical characteristics. These clinical characteristics were only recorded for the index visit. The medication group is derived from a medication list which extends across multiple fields.

As part of an analysis plan for a manuscript the investigators proposed to perform a “conditional binomial test” comparing the two drug groups, the rationale being to test whether or not group A drugs are more likely to be associated with itching than group B drugs. The staff statistician was concerned that this inference is unsound but wasn’t able to communicate why to the satisfaction of the investigator. Additional questions included what to do with the multiple and variable number of records per patient, as well as the pattern of systematically missing data. A planning meeting was scheduled.

In advance of the planning meeting the collaborative statistician sent the investigator an email illustrating the concern about comparing the drug categories. We don’t have medication lists for the 20,000 (i.e., 20,500-500) records with low itching scores. Those numbers could turn out to be 18,000 records from group A, 300 of which indicated severe itching (1.7% incidence rate) and 2,000 records from group B, 200 of which indicated severe itching (10% incidence rate). On the other hand, the conditional binomial test would notice that of the 500 data points, 300 (i.e., 60%) are associated with group A and 200 (i.e., 40%) are associated with group B, and the inaccurate

conclusion would be that group A drugs are more likely to be associated with severe itching, even though the incidence is actually much lower than that of group B. In the absence of an appropriate denominator, in turn caused by the fact that the medication list is missing for those visits with itching scores of 0-3, it isn't possible to say that group A is more likely to be associated with itching than group B. The investigator accepted this argument, and asked that the meeting discuss what can be demonstrated instead.

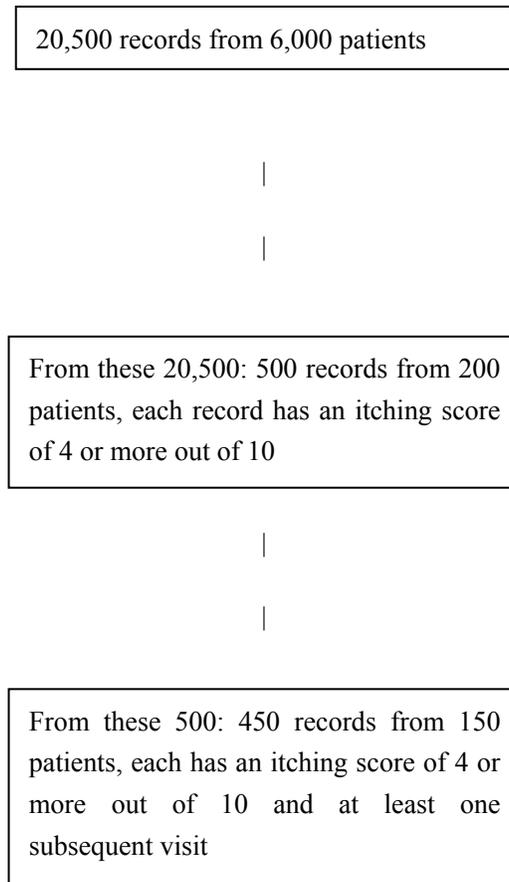


Figure 2. Data Flow

Figure 2 illustrates the flow of data in case study 2.

Table 2. Data Array for First 2 Patients (n=500)

Patient	Visit	Index visit	Gender (and other clinical characteristics)	Medication group	Itching score	Discussion of itching	Action taken	Itching score at next visit
2	2	Yes	M	A	4	N	N	5
2	3	No		A	5	N	N	3
3	8	Yes	F	B	9	N	N	.

In response to the questions “what story do you want to tell?” and “which elements of this story are the most important?”, it turned out that what the investigator wanted to do is to demonstrate that for this group of patients severe itching isn't given adequate attention. One piece of the desired argument was that clinicians should consider the possibility of severe itching in all patients, not just those receiving drugs from group A, and that for this purpose it's sufficient to note that some of the cases of severe itching aren't associated with group A drugs, and also that a formal statistical test isn't required. The analysis plan below was recommended, populated with illustrative data and the conclusions that can be drawn from that data. The investigator agreed that the analysis plan is appropriate, and

the staff statistician implemented it.

For the present purposes the analysis plan (Table 3) will be described in terms of tables, to assist in identifying its inputs, although in the actual manuscript some of these tables might be replaced with figures.

Table 3. Analysis Plan

Characteristics of the patients with severe itching (n=200 patients)

Gender	Frequency
Male	Xx
Female	Xx
Next characteristic...	

Conclusion: Discuss any interesting features. These are the general characteristics of the 200 patients with at least one record of severe itching, not the 500 episodes of severe itching, and should be interpreted accordingly. Patient characteristics were recorded at the index visit – that is, the first visit with severe itching.

Distribution of itching scores (n=20,500 observations)

Score	Frequency
0	15,000
1	3,000
2	1,000
3	1,000 (0-3 adds to 20,000)
4	100
5	100
6	100
7	50
8	50
9	50
10	50 (4-10 adds to 500)

Conclusion: Itching is relatively uncommon (incidence 500/20,500 visits), but is nevertheless common enough to merit clinical attention. For patients with itching scores in the clinically worrisome range of 4-10, a notable proportion of these scores are very high. This information might be presented as a histogram.

Distribution of clinician discussion of itching (n=500 observations with severe itching)

Clinician assessment	Frequency
Yes	50 (10%)
No	450

Conclusion: Recognizing that these figures might be different for the n=20,000 records without additional data collection (but having no reason to believe that they will be), clinicians seldom discuss itching with their patients.

Distribution of clinician action in response to severe itching (n=500 observations with severe itching)

Clinician assessment	Frequency
Yes	20 (4%)
No	480

Conclusion: Clinicians seldom take action in response to severe itching, likely because they don't ask in the first place.

Distribution of next itching score (n=450 observations from patients with severe itching and a subsequent record in the database)

Next score	Frequency
0-3	225
4-10	225 (50%)

Conclusion: In approximately half of the cases severe itching didn't resolve on its own. It would have been better to take action during the previous visit. (Note: by using the itching score from the next visit complications with different numbers of visits are avoided.)

Table 3 illustrates the analysis plan for case study 2.

Taken as a whole, the above analysis plan tells the story that clinicians don't typically attend to the problem of severe itching, and the result is unnecessary misery, and thus that clinicians should ask about this potential symptom, and to do so for all patients (not just those receiving drugs from group A).

3.2.2 Deconstructing the Interaction

The list below summarizes some of the elements of statistical knowledge which the consulting statistician applied. These elements were combined into framing, and then ultimately solving, the problems at hand.

- It's useful to summarize the study design, often with a picture.
- It's useful to give an illustration of the data array.
- A simple example is often helpful to illustrate a concept.
- To be valid, an incidence rate requires an appropriate denominator.
- Definition of a conditional binomial test, including its purpose and its appropriate use.
- If an investigator suggests a specific statistical procedure they might not necessarily be correct. Review its appropriateness.
- An analysis plan should tell a story. Often the difference between a manuscript that's accepted for publication and one which isn't is whether it tells the right story (and how important the story is). When communicating with an investigator make the analysis plan as concrete as possible.
- For clarity of communication, "show and tell" works much better than "tell".
- Link each of the outputs of the analysis plan to its intended message.
- This type of longitudinal data can be extremely messy – if possible, collapse the data set into one record per person (e.g., as was done in the analysis of the next itching score).

3.3 Definition of Sub-Constructs

In deconstructing the above interactions (which happened to occur on the same day and are quite typical of the professional life of a collaborative statistician), the "information / knowledge" used by the collaborating statistician can be placed in various categories.

The first category is *specific and concrete information about statistical procedures* – for example: the technical mechanics of a power calculation assuming that its inputs are known. Such information falls within the technical and analytical competency, and is taught either through a formula or, in more advanced training, through the statistical distributions and the logic that underpins the formula.

Power calculations (i.e., to justify the choice of sample size) are critical to grant applications, and in standard cases such as a 2-group design they can be performed by virtually anyone by following the directions from a website. Nevertheless, understanding the assumptions and the underlying mathematics of a power calculation is beneficial as it allows the statistician to identify when standard formulae are inappropriate, and also to suggest what to do in response. Here, for example, the means and standard deviations for the two groups – that is, the information for which a website would prompt – were unavailable, but it turned out that an effect size could be used instead.

Another example of specific and concrete information about statistical procedures pertained to the conditional binomial test, a test that is relatively common although not always included in a first-year statistics curriculum. In practice, details about the test are something that could be looked up online, and one of the lessons for effective collaborative practice is simply that it's best to treat investigator recommendations about statistical procedures as

provisional. For completeness, what the conditional binomial test actually does is to test the null hypothesis that “among cases with severe itching, exactly 50% came from group A”, a hypothesis that isn’t relevant to the investigator’s story. Having a general understanding of the mathematical logic behind the conditional binomial test would help provide a way forward.

Another category of information is *specific to the biomedical content being studied*. In case study 1, it was helpful to know that effect sizes for apps typically range from small to moderate, and also the sample sizes for typical randomized trials in this area. Having this information at hand allowed for a recommendation to be made in real time during the consulting session, and is an example of the knowledge that’s accumulated over time, but it isn’t absolutely essential. In particular, a statistician without experience in this area could have asked the investigator about the size of typical trials and also to recommend some articles describing typical trials, from which effect sizes could have been derived. Case study 2 didn’t use any outside information about the biomedical content being studied.

The main pedagogic issue pertaining to teaching biomedical content to students is its vastness – for example, statisticians in training are unlikely to know whether their collaborators will be studying cancer pain, lung function, infectious diseases, genetics, and so forth. Our program’s response to this challenge was to develop a general survey course on biology and medicine, for an audience of statisticians, assuming no previous training in biology. The course focuses on basic terminology and general principles, and much of the point of the exercises isn’t necessarily the specific information that’s gleaned but instead is practice for life-long professional learning. Not at all a formal evaluation, but I personally found that sitting in on the course has increased my effectiveness as a collaborative statistician. (And, also, I now know that the answer to the “how” in biology and medicine is usually “it depends” and the answer to the “why” is usually “evolution”.)

Another category of information is *general knowledge about biomedical studies, especially study design*. Considering case study 1, the most basic information would include the differences between randomized and non-randomized designs, the differences between randomizing at the level of the individual versus the cluster, etc., all of which would typically be covered in a class in clinical epidemiology, among others. Such a class would also probably cover the strengths and weaknesses of those designs on a general level, and perhaps even some basic rubrics for selecting one design over another. Making those rubrics explicit and supplementing them with case studies seems critical, as the question of “how do I go about deciding on a study design?” or, perhaps, “how should I organize my thinking around how to decide on a study design?”, is a topic which students of statistics often find to be particularly challenging.

In addition to the above, over time collaborative statisticians also develop tacit knowledge in the form of “in this situation this is the most important consideration in selecting a design”. One exercise that can potentially assist in progressing towards this goal is to ask mentors and colleagues questions such as “what alternatives did you consider in designing this study and why did you ultimately choose as you did?” Another technique is to do the explicitly do the same in one’s own work and then request feedback. One way to simulate this in a classroom setting is through case study exercises where alternative designs are presented and the student is given the task of describing the strengths and weaknesses of each, perhaps as the background for a debate on the topic of which design is best in that particular circumstance.

The power calculation in case study 1 illustrated at least two elements of tacit knowledge. The first element, and one that can be simply stated and also practiced using case studies, is that the decision making around sample size should almost entirely focus on the primary hypothesis. The rationale for this rule of thumb is that failing to follow it risks a significant amount of rework. For example, a study might be sufficiently powered to test the primary hypothesis but not a secondary one. The investigator might increase the sample size to achieve this latter goal, only to discover that it is logistically infeasible. Then, perhaps, design alternatives might be entertained to address this secondary hypothesis, which then require the analysis plan for the primary hypothesis to be revised. As might be imagined, such scenarios comprise the introduction to more than one horror story among collaborative statisticians. The educational goal here is increase the likelihood that the student experiences such a horror story second-hand rather than first-hand.

Another element of tacit knowledge pertains to assisting the investigator in generating the inputs to the power calculation. Here, a critical piece of information is the rule of thumb that minimum clinically important difference is often better to specify than the likely effect size. This concept can easily be included in classroom instruction, and especially helpful would be exercises that allow the student to practice the various ways that the minimum clinically important difference can be derived. In the case study the minimum clinically important difference was the effect

size that administrators would require in order to be willing to roll out the app.

In case study 2, the general knowledge in question was what constitutes an appropriate denominator for an incidence rate. This is covered in the basics of clinical epidemiology, among others. A possible classroom exercise on this topic would be to consider various scenarios and ask whether what is being estimated is incidence or prevalence, and also whether the proposed inference is sound. In this portion of the case study what was particularly helpful for the investigator was translating this basic concept of clinical epidemiology into a simple numerical example, which touches on the communication competency. More generally, one of the principles of effective communication is that a simple example is often an effective way to illustrate a concept.

In case study 1, the information that, across large areas of medicine, research starts with an efficacy study which compares the best bet against the weakest possible comparator, and only if positive proceeds to subtler distinctions, is another type of general observation about study design. This is an observation – one which is more about “what usually is” than “what should be” -- that would be made from reading the literature (i.e., to observe what’s published) and/or participating in the deliberations of a study section (i.e., to observe what’s funded). The discussion around how to select an appropriate control group in a randomized trial also falls into this category. Most collaborative statisticians gradually glean this information over time, with internship being a natural place to start. The careful study of review articles can facilitate the learning process.

Another category pertains to *insights about the process of effective collaboration* gained over time. For example: (a) don’t assume that the question an investigator poses to the statistician is the correct one; (b) begin collaborations by reviewing the perceived rationale for the study design; (c) create an explicit statement summarizing the rationale for the study design; (d) identify the study question and also why that question is important; (e) summarize the study design with a picture if possible; (f) to facilitate communication about data structures, illustrate the data array; (g) the analysis plan should tell a story; (h) explicitly discuss what story is best to tell; (i) show and tell works better than tell; and (j) link each of the outputs of the analysis plan to an explicit message.

One way to help speed the process by which students become effective collaborators is to supply them with standard rubrics and reporting forms. For example, in case study 1 the narrative description of the investigator’s grant includes (among others) some general background, the study question and the proposed study design, and students can practice extracting information provided by investigators into such a format. In case study 2 the reporting forms included a figure summarizing the data flow and a concrete illustration of the data array (with commentary). Students can similarly practice feeding information back to instructors in these types of formats. Moreover, such practice can eventually extend from summarizing the information from written documents to doing the same within simulated interactions with investigators (or supervised interactions with real investigators). Essentially what is being practiced is the “mirroring” exercise associated with active listening where the statistician tells the investigator “this is what I understand you have told me”, with “this is what I understand you have told me” not reported verbatim but, instead, translated into a format which is useful to the statistician but nevertheless understandable to the investigator as well.

A final category of information is *high-level synthesis*. One example of a high-level synthesis is that, in a randomized trial, it’s usually more efficient to compare two groups rather than three. Another example is that case study 2 doesn’t contain the sort of data that’s typically amenable to specialized techniques of longitudinal analysis. (Demonstrating why these statements typically hold true isn’t relevant here.) High-level synthesis might not be a realistic goal for students who are simultaneously learning the basics of a discipline but, in the spirit of constructivism, its eventual attainment can be assisted by having instructors describe not just what they would do in a particular situation, but also why. In effect, this serves to explicitly describe their mental maps of the discipline, and helps to place the more specific information that the student is learning into better context. Indeed, one of the most consistent comments from teaching a course in statistical modeling to physicians (Samsa *et al*, 2012) was a request to not just focus on the “how to do it” but also on the “when to do it and when to do something else instead”, which in essence is asking for a mental map.

4. Discussion

The four core constructs for effective collaborative statistics can be defined as technical / analytical competence, substance-matter knowledge, communication, and problem-framing / problem-solving. One hypothesis was that these competencies can be separated into sub-competencies and, in fact, our case studies were able to identify such sub-competencies (i.e., specific and concrete information about statistical procedures; information specific to the biomedical content being studied; general knowledge about biomedical studies, especially study design; insights about the process about the process of effective collaboration; and high-level synthesis), which in turn could be

mapped back into the above four core constructs.

Technical / analytical competence is the current focus of statistical education, and the other competencies are addressed more haphazardly and with less emphasis. We hypothesized that one of the main reasons that these other three competencies receive less attention is that they are less straightforward to teach than the technical / analytical competency. While not directly testing this hypothesis, what our case studies did suggest was that the sub-competencies for substance-matter knowledge, communication and problem-framing / problem-solving are at a higher level of abstraction than the technical / analytical sub-competencies, and this is consistent with the conclusion that they are less straightforward to teach.

Considering the question of “why the disconnect between what is important and what is taught”, we propose that the reasons that statistics programs so strongly focus on the technical / analytical competency are multiple. One reason is probably inertia: this is the way we were taught as students and so becomes the default approach for busy faculty members whose plates contain many other things in addition to teaching. Not to mention that the technical and analytical information is unquestionably important – it is simply that for collaborative statistics other competencies are vitally important as well.

A barrier to addressing the communication and problem related competencies is that some instructors appear to consider these to be fixed traits – as in “you are either born to be an effective communicator or you aren’t” – rather than competencies. While it’s true that some people have more talent for communication than others (and similarly for problem-solving), it’s also true that what’s at issue isn’t the construct of communication as a whole, but rather a small sub-construct pertaining to communication around various statistical issues. Specific techniques can be very helpful – for example, creating reports such as Tables 1 and 2 to describe the data should facilitate communication about data structures, practice in the technique of mirroring should facilitate communication with investigators about various issues, etc. A combination of sharing specific techniques with students, plus an emphasis in “communicating across the curriculum” (i.e., analogous to “writing across the curriculum”, but with students also given opportunities to present orally in various contexts) ought to go a long way towards meeting realistic goals.

A final barrier is the relatively lack of scholarship on the question of how to divide potentially nebulous competencies into smaller units that are amenable to instruction and assessment. For example, in the absence of a description of the sub-constructs that the overall competency of communication includes it’s easy to design a standard class in statistical consultation and just hope for the best. Indeed, a typical reaction to the notion of elevating the competencies other than the technical / analytical to a more central place in the curriculum has been “that’s a great idea, and that’s the stuff that students need to learn, but how in the world do you plan to do it?”

Briefly, and without intending this to be either a program evaluation or even a detailed description, our program’s response to the challenge of addressing these additional competencies has been to (a) develop a course in biology and medicine for statisticians intended to introduce basic concepts, basic terminology, and a framework which could facilitate subsequent independent learning; (b) develop a course in statistical practice that’s analogous to the statistical consulting courses offered elsewhere but with the addition of specific rubrics that can be of practical assistance; (c) embed the analysis of real data from actual research projects into classroom instruction and make this part of communicating across the curriculum; and (d) develop an extensive internship program (i.e., recognizing that individual mentorship around a real application is the ideal way to practice higher-level skills, and also recognizing that even though mentorship is a high-level skill that can’t be reduced to a checklist, its implementation can be made more systematic).

A final hypothesis was that identifying sub-competencies can help provide a way forward in curriculum design. Although we have suggested how various exercises might address these sub-competencies, the degree to which these sub-competencies can assist in curriculum design remains an open question.

This deconstruction of the work of a collaborative biostatistician is qualitative and provisional. From the perspective of the scholarship of determining how to teach communication, problem framing and problem solving within the context of collaborative biostatistics what is critically needed is a finer-grained, evidence-based and fully-actionable description of what these competencies actually entail.

Competing interests

The author declares that he has no competing interests.

References

- Baskerville, J. C. (1981). A systematic study of the consulting literature as an integral part of applied training in statistics. *The American Statistician*, 35(3), 121-123.
- Begg, M. D., & Vaughan, R.D. (2011). Are biostatistics students prepared to succeed in the era of interdisciplinary science?: (and how will we know?). *The American Statistician*, 65(2), 71-79. <https://doi.org/10.1198/tast.2011.10222>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum.
- McCulloch, C.E., Boroto, D.R., Meeter, D., Pollard, R., & Zahn, D.A. (1985). An expanded approach to educating statistical consultants. *The American Statistician*, 39(4), 159-167.
- Samsa, G.P., Thomas, L., Lee, L.S., & Neal, E.M. (2012). An active learning approach to teach advanced multi-predictor modeling to clinicians. *Journal of Statistics Education*, 20(1), 1-31. <https://doi.org/10.1080/10691898.2012.11889631>
- Sharples, S.J., Yeend, E., Francis, B., & Booth, J. (2010). Developing statistical consultancy skills in post-graduate students: a case study. In: *ICOTS8 Proceedings*. International Association for Statistical Education, Auckland, New Zealand.
- Taplin, R.H. (2003). Teaching statistical consulting before statistical methodology. *Australian and New Zealand Journal of Statistics*, 45, 141-152. <https://doi.org/10.1111/1467-842X.00270>
- Taplin, R.H. (2007). Enhancing statistical education by using role-plays of consultations. *Journal of the Royal Statistical Society Series A*, 170(2), 267-289. <https://doi.org/10.1111/j.1467-985X.2007.00463.x>
- Van Belle, G. (1982). Some aspects of teaching biostatistical consulting. In J.S. Rustagi & D.A. Wolfe (Eds.), *Teaching of statistics and statistical consulting* (pp.343-365). New York: Academic Press. <https://doi.org/10.1016/B978-0-12-604540-6.50035-8>
- Winkvist, K. (1990). Statistics in Action: Learning and Comprehending by Doing. In *ICOTS-3 Proceedings*. Retrieved from www.stat.auckland.ac.nz/~iase/publications/19/book2/c4-9.pdf