Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

IMPROVING PRIMARY CARE ACCESS USING SIMULATION OPTIMIZATION

Hari Balasubramanian
Ritesh Banerjee
Melissa Gregg

Division of Health Care Policy and Research
Mayo Clinic
Rochester, MN 55905, U.S.A.

Brian T. Denton

Edward P. Fitts Department of
Industrial & Systems Engineering
North Carolina State University
Raleigh, NC 27695-7906

ABSTRACT

Primary care providers (PCPs) provide the majority of care patients receive during their lifetime. We consider the problem of determining the size and composition of physician panels in primary care. A physician's panel consists of a set of patients and each patient belongs to one of many different health-related categories. Using real data collected at the Mayo Clinic at Rochester, we propose a multi-period metaheuristic simulation optimization model for determining the panel design of a set of physicians working in a primary care environment. The model seeks to maximize patient visits to their own providers, reduce waiting times, and minimize overage.

1 INTRODUCTION

Primary care providers (PCPs) are typically the first point of contact between patients and health systems. From a patient's perspective, PCPs provide the majority of care they receive during their lifetime and are responsible for a variety of health services including preventive medicine, patient education, routine physical exams, and the coordination of complex episodes in which patients are referred to medical specialties for secondary and tertiary care. Despite primary care's crucial role, the Institute for Healthcare Improvement (www.ihi.org) reports that 40% of emergency department visits are not urgent, rather, they take place because patients cannot see their PCP. In addition, from 1997 to 2001 the percentage of people reporting an inablity to obtain a timely appointment rose from 23% to 33% and in 2001, 43% of adults reporting an urgent condition were unable to receive care when they wanted.

There are many benefits to regularly seeing one primary care provider, most of which are associated with improved long term quality of care. Gill, Mainous, and Nsereko (2000) point to several studies which show that patients who regularly see their own providers are

1. more satisfied with their care,
2. more likely to take medications correctly,
3. more likely to have problems correctly identified by their physician, and
4. less likely to be hospitalized (for other studies see, Christakis et al. (2000), Becker, Drachman, and Kirscht (1974) and Gill and Mainous (2000)).

We are therefore faced with two competing objectives when thinking about questions of primary care access: should we minimize waiting times at the cost of continuity of care or ensure continuity but make patients wait?

The access measures of patient waiting times and continuity of care are linked with in the concept of a physician panel. A panel is a set of patients assigned to a physician. Green, Savin, and Murray (2007) investigate under the paradigm of *open* or *advanced access*, the right panel size for a physician. In advanced access (proposed by Murray and Berwick (2003) and Murray et al. (2003)), patients are allowed to book same-day appointments. The underlying assumption is that if capacity is well matched with demand, patients can be served the same day they request an appointment. Green, Savin, and Murray (2007) state that the size of a physician's panel "is the major determinant and the prime lever for achieving the balance between supply and demand." They propose a six step process using a probability model that calculates the amount of overflow (or overtime) for the physician in a given week. They include a brief discussion of how this research applies to a group practice.

Advanced access is likely to work well in situations where the variance of demand for appointments is low or easily predictable. When demand is stochastic, an important open question exists about the design of an access system that balances the competing dual objectives of quick access and continuity. Our focus in this study is to approach this question by investigating *optimal panel design*, by which we mean choosing the "best" combination of composition and size of a set of physician panels in a group practice. In

other words, what should be the number of patients and set of patient characteristics in each physician's panel? Some patients tend to have a higher need to see their provider. For instance, elderly male patients will have a different request rate than female patients in the 20-30 age range. Panel design thus affects the volume and variation in patient appointment requests and in turn provider availability.

We propose a simulation-optimization approach to optimize panel design in a group practice to improve patient access to primary care. We use data collected from the Mayo Clinic Primary Care Practice to populate our model. Our model is a weekly aggregation of appointment requests and physician capacities. It is based on the traditional appointment scheduling model where some patient requests are placed on a waiting list until an appointment becomes available, and others are seen in the same period. Our model seeks to optimize panel design to meet the following objectives:

1. minimize average time a patient waits to see a provider,
2. maximize a patient's visits to his/her PCP, or minimize the average number of redirections to other providers, and
3. minimize the amount of additional capacity (overage) added in every period.

2 THE PANEL DESIGN PROBLEM

Patients fall into one of many health-related categories. Factors such as age, health status, geographic location, patterns of historical appointments (appointment types, urgent vs. non-urgent requests, etc.) can be used to define a classification of patient types. The purpose of such a classification is to reflect the reality that patient request rates from the different categories may vary significantly. We use twenty eight categories based on age and gender starting with separate groups for males and females aged 18-23,and with the rest of the categories being in five-year increments. The last two groups are patients aged over 83. While other elaborate categorizations are possible, the age-gender categorization gives a simple, intuitive classification that can be tested in our preliminary model.

The panel design problem is an allocation or assignment problem. Given a set of health-related categories, and a set of physician panels in a group practice, how many patients from each category should be assigned to each panel?

Consider a 1-period example with 3 patient categories and 3 physician panels as shown in Figure 1. The three patient categories have sizes of 22, 37 and 45 patients. Assume that all of these patients request for an appointment in the period. Each physician can see a maximum of 35 patients the period. The number of patients assigned from a category to a panel are indicated on the arrows in the figure. These values (later called x_{ij}) determine the panel design for this example. These values are also the decision variables in our study: our simulation-optimization approach will seek to optimize these values.

The dashed lines indicate redirection of flow for the single-period model. Redirections mean that patients have to see another provider since their PCP is unavailable. Clearly, for the 1-period deterministic problem, the panel design is not optimal. If the number of patients assigned from category 3 to panels 1, 2 and 3 is changed to 20, 12 and 13, all patients will see their own PCP and there will no redirections.

If we add another period, redirections may either be to the same provider in the next period or to a different provider in the same period. These are shown using solid arrows in the figure. Redirections in this situation depend on whether patients are willing to see a different provider in the same period or are willing to wait for their own provider in the next period.

The complexity of the model increases significantly when we assume a multi-period model (multiple days or weeks) where in every period, the number of appointment requests for each panel is a random variable and inter-period variability is significant due to seasonality and temporal effects. Depending on the panel design, physicians could end up having excess capacity in one period and having an over-full calendar in another, affecting access metrics. Optimizing the composition and size of panels thus becomes important.

2.1 Simulation Model and Notation

We now describe our simulation model and the performance measures by which a panel design solution is evaluated. We consider our model as a first step in the development of a more realistic representation of primary care access. We make the following key assumptions that we plan to relax in our future research efforts:

1. The patients who are redirected can be seen by any of the physicians, irrespective of their ailments. In practice, certain sets of patients are better suited to certain physicians. Physicians may also have certain preferences in this regard.
2. While we consider FTE (Full Time Equivalent) information of the physicians to determine capacities, we assume that these capacities do not change from week to week. In practice, capacities are set in anticipation of how the time of the year might affect patient request rates.
3. The categorization used for patients in our model is based on age and gender. The categories are important in that they determine the request rates of their patients: we assume that patients within a category behave in fairly similar fashion.

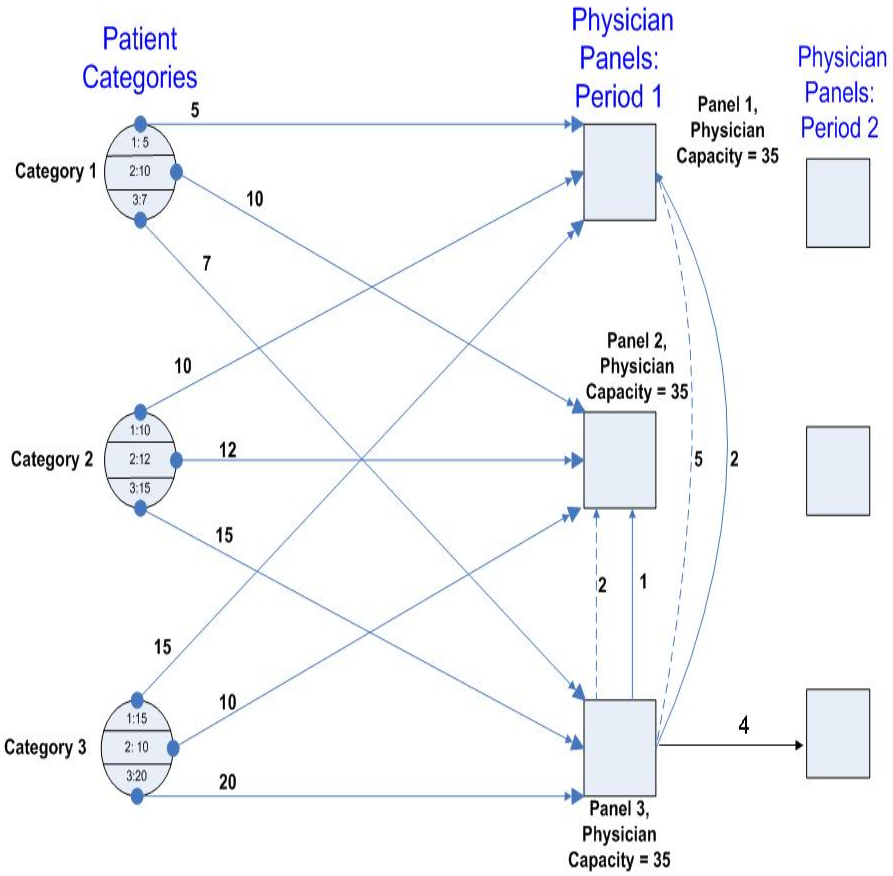

Figure 1: Simple single and two-period panel design example

We now proceed with describing our model. Let m be the total number of patient categories and n be the total number of physician panels in a primary care environment. Let x_{ij} represent the number of patients assigned from category i to panel j. The vector $X = x_{ij}$ values thus describes the *panel design*. Note that this is treated as a one time decision in our model.

Let S_i represent the total number of patients in category i. Thus $\sum_{j=1}^{n} x_{ij} = S_i \forall i$. Let T be the total number of periods in the model, and $d_{ik}(\omega)$ be the number of appointments requested from category i in period k under scenario ω. The number of requests to a particular physician's panel from a given category is given by $f_{ik}(\omega)x_{ij}$, where $f_{ik} = d_{ik}(\omega)/S_i$. In other words, f_{ik} is the fraction of patients in category i who request an appointment in period k under scenario ω. We also assume that each of the m physicians can see at most C_j appointments in every period.

For a given set of x_{ij} values, our simulation model works as follows: physician calenders start empty (no appointments) in the first period and move from period to period. In each period, patients request appointments from their PCPs. Patient requests in each period are satisfied on a first-come-first-served basis: requests originating in earlier periods are filled first. Ties between requests arising in the same period to a given physician panel j are broken arbitrarily. If the capacity C_j for a particular panel is exhausted, patients have one of two options: to see another provider within the same period or to wait for a future period to see their PCP. We assume that a fixed proportion of patients are willing to see other providers in the same period. These requests are filled by physicians who have available capacity. Let $L_k(\omega)$ be the sum of unused capacity over all physicians in period k under each scenario.

Let $b_k(\omega)$ represent the total number of patients that were unable to see their provider upto period k under scenario ω. Let $g_k(\omega), 0 \leq g_k(\omega) \leq 1$ be the proportion of these patients who are willing to see other providers in the same period. Then $G_k(\omega) = g_k b_k$ requests have to be accommodated in the same period, while the rest are assigned to a future period. We repeat this procedure for every period $k = 1, \ldots, T$.

We now give expressions for two of our performance measures:

1. $R_k(\omega)$ which denotes the number of redirections to other physicians in period k, and
2. $A_k(\omega)$, the additional capacity or overage for period k.

If $G_k <= L_k$, then

$$R_k = g_k b_k \tag{1}$$

$$A_k = 0. \tag{2}$$

And if $G_k > L_k$,then

$$R_k = L_k. \tag{3}$$

$$A_k = g_k b_k - L_k. \tag{4}$$

We also give the expression for average waiting time of patients. Let $N_l(\omega)$ be the number of appointment requests that are filled l weeks in the future. Then the average waiting time $W(\omega)$ of patients under scenario ω is given by:

$$\frac{\sum_{l=1}^{T} lN_l}{\sum_{l=1}^{T} N_l} \tag{5}$$

Thus our optimization problem is to find an X that minimizes the following expectation:

$$E_\omega \left\{ c^A \sum_{k=1}^{T} A_k(\omega) + c^R \sum_{k=1}^{T} R_k(\omega) + c^W W(\omega) \right\} \tag{6}$$

Where c^A, c^R and c^W are costs for additional capacity or overage, number of redirections and waiting time.

To summarize, our multi-period simulation model enables us to determine measures such as average waiting time, average number of redirections and average additional capacity for given panel design. In Section 3, we describe a metaheuristic approach evaluate and optimize various panel designs using the simulation model.

3 A GENETIC ALGORITHM FOR OPTIMIZING PANEL DESIGN

We now propose a genetic algorithm designed to optimize panel design. Genetic algorithms (GAs) have been used extensively over the last two decades for deterministic combinatorial optimization problems. With recent increases in computing power they are becoming increasingly useful in simulation optimization approaches. GAs allow for evaluation of a large number of possibilities and also for an intelligent search to determine the solution with the best expected value of the objective under consideration.

Our panel design genetic algorithm (PDGA) searches to panel design solution space and uses the simulation model proposed in Section 2.1 to determine the quality of each solution. The main steps of PDGA are listed below:

1. *Encoding*: The encoding of each solution in PDGA is simply a vector of x_{ij} values i.e the vector X. Since there are m categories and n panels, the vector is mn long and indicates the number of patients assigned from each category to each panel.
2. *Initial Population*: The initial population of PDGA is randomly generated. The portion of a certain category that will assigned to a panel is randomly determined, but meets the following feasibility constraint for each category: $\sum_{j=1}^{n} x_{ij} = S_i, \forall i$. We also insert into the initial population the panel design currently in use at the Mayo Clinic primary care practice. The size of the population is N_P, and is kept constant for all generations of the PDGA.
3. *Evaluation and Fitness*: Each panel design solution in the population is evaluated using the simulation model proposed in Section 2.1. The fitness of each solution is the objective function stated in 2.1. We assume in this study that the costs for redirections, overage and waiting time are the same. We use 50 replications of the simulation to determine expectations on our performance measures and to compare solutions.
4. *Crossover*: The crossover operation exchanges panel-category assignments across two parents to produce two offspring. Figure 2 illustrates the details of these exchanges using an example.
   The number of such exchanges between parents is set to a prefixed value. In total $N_P/2$ crossovers are carried out to produce N_P offspring.
5. *Mutation*: The N_P offspring undergo mutation based on a pre-specified mutation probability. Figure 3 illustrates mutation with an example. Note how panel assignments within a certain categories are exchanged.
6. *Selection to the next generation*: At the end of mutation, there are N_P solutions in the original population and N_P offspring. Of these $2 * N_P$ solutions, $N_P/2$ of the best solutions are chosen to the next generation, while the remaining $N_P/2$ are chosen randomly among those not already picked.

4 EXPERIMENTAL RESULTS

Our scenarios for simulating and evaluating each panel design solutions in PDGA are generated randomly using sampling with replacement. Each period in our model

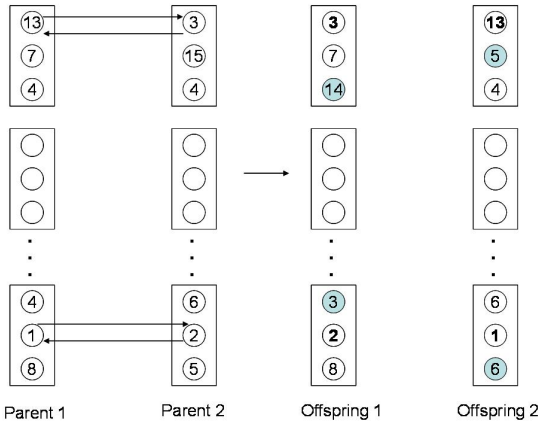

Figure 2: Illustration of crossover. The rectangles represent categories while the circles indicate panels. The number in the circles indicate the number of patients assigned from a category to a panel. Parents exchange several of their assignments (randomly chosen) to produce new offspring. Note that in order to maintain category sizes, other category-panel assignments in the offspring (again randomly chosen) are adjusted (in the offspring, the panels corresponding to these changed numbers are shaded).

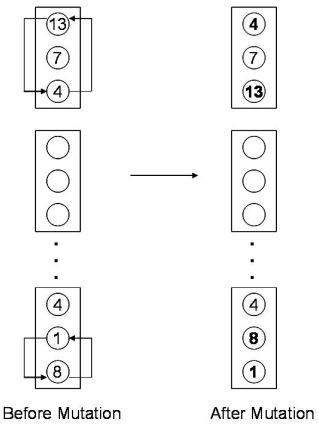

Figure 3: Illustration of the mutation. Panel assignments *within* certain categories are exchanged.

corresponds to a week, and $T = 52$ weeks. The request rates from each category are sampled from weekly visit data at the Mayo Clinic Primary Care practice. We use data collected for weeks from 2004-2006, which gives 150 weeks of information. Our sampling method is selective: for any T we sample from weeks $max(T-4,0), min(T+4,52)$ from years 2004-2006. This allows us to express time-of-year effects for patient appointment requests.

Our model also corresponds to size of the Mayo Clinic Primary Care practice, which consists of 39 physician panels. We consider 28 categories based on age and sex (14 age categories and 2 gender categories). To determine capacities, we use FTE (Full Time Equivalent) information of the physicians i.e. the percentage of time spent by these physicians in primary care. The capacity estimate for a physician is his/her FTE value multiplied with the upper bound on the number of patients that a physician can see in a week (5 8-hour days).

As stated before each panel design solution in the PDGA was evaluated using 50 replications of the simulation described in 2.1. Data from the first 15 weeks was truncated to remove the initial bias. PDGA had a population size of 25 and was terminated after 50 generations. The number of exchanges for each crossover operation (the exchanges are shown in Figure 2) are fixed at 10, while the number of exchanges for a solution undergoing mutation is fixed at 8. We arrived at these parameter settings using trial and error, based on pilot runs and with computation time considerations.

We now compare the best solution obtained using PDGA to the current panel design used at the Mayo Clinic primary care practice (see Table 1). The wait time is in weeks and is the average per patient; redirections indicate the average number of patients sent to other providers on a weekly basis; and overage indicates the average additional appointment slots that had to be added on a weekly basis. See Section 2.1 for a detailed explanation of these performance measures. It is clear that given the current assumptions, that PDGA outperforms the current panel design. The differences were statistically significant. This was verified by building a 95 percent confidence interval of the differences in the performance measures; for all three of the objectives, the CI contained 0.

Figure 4 shows graphically the differences in wait times involved in the two solutions. The x axis indicates the the number of weeks. The y axis indicates the average number of patients who waited i weeks or more, where i is the corresponding value on the x axis. The PDGA solution performs significantly better but the differences taper off as the number of weeks increases.

In summary it appears from the preliminary results that the PDGA shows promise. But it needs to be noted that the current panel design is based on various factors - such patient and physician preferences - and that our model does not consider. As noted before, several of the assumptions of the model do reflect what happens in practice.

5 CONCLUSIONS AND FUTURE RESEARCH

We propose a multi-period simulation-optimization model to optimize panel design and improve access to primary care. The objectives of our model are to minimize the expected number of redirections, waiting time and overage. Our GA approach demonstrates that significant improvements over the current panel design are possible. However, it is

Table 1: Table listing the performance measures of the currently used panel design and the PDGA model.

	Wait Time	Redirections	Overage
	Current Panel Design		
Mean	6.82	531.10	31.48
Std.Dev	0.21	26.43	20.78
Upper 95% CI	6.87	538.43	37.24
Lower 95 % CI	6.76	523.78	25.72
	PDGA Solution		
Mean	4.50	230.13	0.09
Std.Dev	0.18	26.77	0.34
Upper 95% CI	4.55	237.55	0.18
Lower 95 % CI	4.44	222.71	0.00

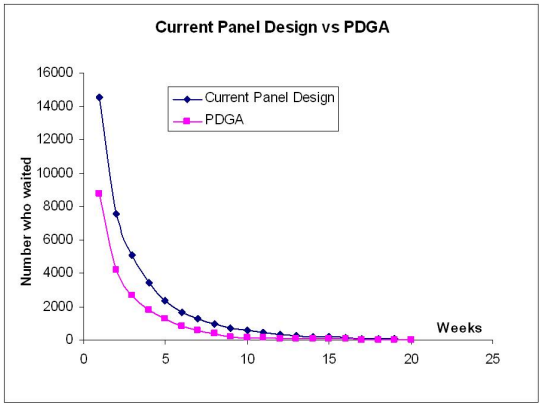

Figure 4: Graphical comparison of current panel design and PDGA panel design.

necessary to consider the practical implications of altering the current panel design. Rather than suddenly change the PCPs of many patients, the fact that patients leave the system (owing to death and other causes) and new patients enter the system on a regular basis can be leveraged to ensure that changes in panel design smoothly implemented.

In addition several key assumptions in our approach need to be relaxed to obtain a more realistic representation of primary care access. This effort is part of our ongoing research. Some of our main focus areas are:

1. Our model currently assumes that physicians have fixed capacities. We plan in the future to use data on weekly capacity.
2. Our patient categorization currently assumes that the age and gender categories proposed are good predictors of patient request rates, and that patients within a given category have a degree of uniformity with regard to request rates. However, an interesting research area would be to determine a categorization that better predicts patient request behavior and hence improves the accuracy of the model.

We also plan to investigate the concept of patient assignments to a small group of providers. This is similar to the concept of "chaining" in manufacturing and represents a balance between the ideal situation where each patient sees his PCP every time she needs care with no waiting, and the reality where often patients see providers other than their own.

ACKNOWLEDGMENTS

We thank Jim Naessens (Consultant at the Mayo Clinic, Rochester) and Sidna Scheitel (Consultant at the Primary Care Inernal Medicine at the Mayo Clinic, Rochester) for their valuable inputs. This research was funded in part by the Mayo Small Grants Program (SGP).

REFERENCES

Becker, M., R. Drachman, and J. Kirscht. 1974. Continuity of pediatrician: New support for an old shibboleth. *Journal of Pediatrics* 84:599–605.

Christakis, D., et al. 2000. The association between greater continuity of care and timely measles-mumps-rubelle vaccination. *American Journal of Public Health* 90:962–965.

Gill, J., and A. Mainous. 2000. The role of provider continuity in preventing hospitalizations. *Archives of Family Medicine* 7:352–357.

Gill, J., A. Mainous, and M. Nsereko. 2000. The effect of continuity of care on emergency department use. *Archives of family medicine* 9:333–338.

Green, L., S. Savin, and M. Murray. 2007. Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety* 33:211–218.

Murray, M., and D. Berwick. 2003. Advanced access: Reducing waiting and delays in primary care. *Journal of the American Medical Association*.

Murray, M., T. Bodenheimer, D. Rittenhouse, and K. Grumbach. 2003. Improving timely access to primary care: Case studies of the advanced acccess model. *Journal of the American Medical Association* 289:1042–1046.

AUTHOR BIOGRAPHIES

HARI BALASUBRAMANIAN is a post-doctoral research associate at the Division of Health Care Policy and Research at the Mayo Clinic in Rochester, Minnesota. Hari

joined the Mayo Clinic after completing his dissertation in Industrial Engineering at Arizona State University. His research interests include scheduling theory, simulation-optimization and metaheuristics with applications in healthcare, service sector and manufacturing. His email address is ⟨balasubramanian.hari@mayo.edu⟩.

RITESH BANERJEE is an Associate Consultant at the Mayo Clinic in Rochester, MN. He received his Ph.D. in Economics from the University of Wisconsin-Madison in 2005. His research interests include the use of optimization techniques to better allocate scarce resources in health care, including in medical decision making. His email address is ⟨banerjee.ritesh@mayo.edu⟩.

MELISSA GREGG is an Statistical Programmer Analyst at the Mayo Clinic in Rochester, MN. She received her bachelors degree in Mathematics from Luther College. She has recently begun working on operations research problems in healthcare. Her email address is ⟨gregg.melissa@mayo.edu⟩.

BRIAN T. DENTON is an Assistant Professor at North Carolina State University in the Edward P. Fitts Department of Industrial and Systems Engineering. Previously he was a Senior Associate Consultant at Mayo Clinic in the College of Medicine. His primary research interests are in the development and application of operations research methods to health care delivery, and his work relating to surgical scheduling won the IIE Transactions outstanding publication award in 2005. He completed his Ph.D. in Management Science at McMaster University in Hamilton, Ontario, Canada. His email address is ⟨bdenton@ncsu.edu⟩.