GEMMA SERVICES

GENERATING ACTIONABLE EVIDENCE FOR PRACTITIONERS

PETER YORK

Gemma Services—a youth-oriented social services agency that operates a long-term residential psychiatric care program for youth—found that their administrative data system, while extensive, generated little data that could be used by front-line practitioners as they worked directly with youth and families.

First Place for Youth (FPFY) helps youth who have aged out of the child welfare system build the skills they need to make a successful transition to self-sufficiency and responsible adulthood. The leaders of both organizations believed strongly in the need to rigorously evaluate their programs so they could produce the kind of evidence that would advance their programs and practices as well as hold themselves accountable to achieving positive client outcomes.

Like FPFY, Gemma Services considered conducting evaluations using more traditional randomized controlled trials and quasi-experimental evaluation designs. Instead, they chose, as so many providers do, to invest in a program administration data system that would serve to assess, monitor, and evaluate the outcomes of every client throughout their program experience. Both organizations reached the conclusion that, while these data systems served an important program administration purpose, including

being able to report to their funders on their amount of service outputs and costs, they were not meeting their evidence generation needs. This was especially the case when it came to their practitioners and clients. Practitioners were the principal data collectors, spending hours every week gathering information and assessments from clients and inputting this data into the system. However, they received no evidence in return on which they could act to learn about and strengthen their program planning and engagement.

ON-DEMAND EVIDENCE

In 2018, the leaders of both Gemma Services and FPFY sought to build a technological solution that used predictive and prescriptive models to put on-demand actionable evidence in the hands of their practitioners on a daily basis. New analytic studies in justice and child welfare showed it was now possible to do so using the program administration data they already were collecting.

With support from BCT Partners, an evaluation and data science firm with a mission to provide insights about diverse people that lead to equity, the precision analytics (PA) approach Gemma Services adopted is designed to meet different needs than traditional summative evaluation designs, including RCTs and quasi-experimental group comparisons. While such designs are useful in examining the overall effectiveness of a program, practitioners making choices about how to treat or serve the client they are meeting right now need more contextualized and precise information. The PA approach is different in that it applies causal modeling to historical program data by finding similar subgroups of cases and determining the ideal set of tailored program recommendations that maximized their success. Put another way, PA finds naturally occurring experiments where some cases within the same "like" group received a set of services, while others did not, and learns which combination of services maximized success.

Practitioner Buy-In

Gemma Services' leaders knew the success of their adoption of precision analytics would require the buy-in and support of their practitioners. As a first step, they scheduled a series of practitioner meetings to present on the concepts of data science and machine learning and to share how these tools,

combined with their data, could produce case-specific predictive and evidence-based recommendation insights. BCT worked with the practitioners to co-develop ground rules for the precision analytics approach. One key driver for practitioner buy-in was the explicit support of the program leaders and managers that was garnered through the establishment of these ground rules.

Data Readiness

The precision modeling process began with a data readiness audit, which took approximately ten weeks to complete. Both Gemma Services' and FP-FY's data met the minimum requirements identified by BCT: a minimum of 250 cases that do not have too much missing data; at least two years of reliable longitudinal data; case background, situation, history, and needs; program delivery and transactions; ongoing goals, milestones, and accomplishments; and outcomes. Additionally, both organizations requested conducting additional preliminary analyses to test the feasibility of generating results during the full precision modeling process.

The Precision Modeling Process

Once the audit was completed, there were five steps to the precision modeling process:

1. Develop the Analytic Framework

The first step was to develop and refine the program logic model, review and assess the administrative data, and align the logic model with high-level data constructs that were well represented across the administrative data variables.

2. Get the Data Ready

This step, which consumed the largest proportion of the total project time, began with the process of extracting the data, based on the analytics framework, from each organization's data system.

Then, the data had to be transformed into the final set of variables that would go into the precision modeling step. Additionally, different types of statistical and machine learning scaling techniques were used to construct metrics made up of sets of variables that represented measures of all of the components of the program logic model. The transformation process even

included creating "predictive" scales, a technique that builds predictively using machine learning algorithms. This step culminated with selecting the final transformed and logic model–aligned variables for the modeling process and creating prefix tags associated with the different logic model components for each of the variable names to help guide the modeling.

3. Conduct Precision Modeling

The process of building the precision model entailed the evaluator closely collaborating, through multiple screen share sessions, with the data scientists to conduct a series of modeling steps that found matched comparison groups to reduce selection bias, discover what works for each group, and evaluate the effect of what works.

The first step in the precision-modeling process was to *build an "ideal program model."* This model was a predictive model that used all the program dosage, strategy, and goal attainment data to predict the desired outcome.

For example, Gemma Services used their goal attainment data, which represented the tracking of longitudinal changes in thought, behavior, psychiatric, and trauma assessment scores, to predict a child's acuity level at the time of discharge. As noted earlier, Gemma lacked good intervention dosage data; instead, they used assessment score changes as their proxies for what happened to a child during their residential treatment. This first "ideal program model" was able to identify the goal accomplishments most important to reducing a child's behavioral acuity to a level associated with a discharge that was much less likely to return to inpatient hospitalization within the subsequent year. This "ideal model" was approximately 75 percent accurate. Gemma Services' practice expert knew that one size would not fit all children. To reduce selection bias, this goal attainment model produced a probability for every child as to the likelihood they would have achieved a low enough acuity to be considered a success, based on their accomplishment of assessment-based goals.

The next step in the quasi-experimental precision analytics process was to *identify matched comparison groups* based on contextual and baseline intake characteristics that predicted how likely a child was to engage and/or be engaged in the ideal program model.[1] By training machine learning classification algorithms to cluster children into groups based on sharing characteristics that make them equally likely to receive the ideal program

model, this step identified matched comparison groups that could be studied and evaluated during the next steps, thereby minimizing selection bias.[2]

The next step in the precision modeling process is to *train machine learning algorithms* to determine which combination of program interventions, dosages, and/or goal achievements predict the highest likelihood of success for children within each matched comparison group. The evaluator guides the data scientist through an algorithmic training process that identifies the program elements that predict the best outcomes for each matched group. These algorithms are able to produce a ranked and weighted set of program elements that uniquely and in aggregate contribute to achieving the desired outcome. The findings are a group-specific combination of program elements that, when combined, increase the likelihood of a matched cluster of children achieving a positive outcome.

The final step in the precision modeling process is to inferentially *evaluate the effect* the group-specific program model had—in the past—on a matched group of children when some received what works and some, counterfactually, did not. The analytic process included conducting inferential statistical tests (for example, t-tests, ANOVAs, etc.) to determine if those within a specific group who received what works achieved a significantly better outcome score than those in the same group who did not. Effect sizes were also calculated.

4. Automate the Analytic Process

The fourth step in the overall process is to engage the data scientists in automating the analytics workflow such that data extraction, transformation, and loading, precision scoring, and results generation all happen at least once a day without requiring any human to initiate the run. This step also requires determining how to keep the data secure, confidential, and HIPAA compliant throughout the entire data transfer process.

5. Produce Daily Actionable Evidence

This final step serves to design and implement a suite of dashboards for an organization's practitioners, program managers, and leaders to receive on-demand insights—actionable evidence—to be used for case-specific, programmatic, and organizational decision making. This step required data scientists with design and visualization training and experience. Most

organizations now begin with templates created by BCT that use software programs like Tableau and PowerBI to generate dashboards, reports, and visualizations for practitioners.

CHALLENGES AND RESPONSES

Evaluators and practitioners need to be educated about the use of machine learning algorithms and big data analytics for social science research and evaluation. They need this knowledge to develop an understanding of how to collaborate with data scientists to build these tools. BCT worked closely with practitioners from the very start to co-develop "ground rules" for the use of algorithms and to help them understand concepts such as "selection bias."

*Separate individual practitioner performance from the dashboards***.** Most practitioners embraced the opportunity to have data-driven real-time feedback to guide their work and make course corrections. At the same time, practice experts and program managers quickly realized the dashboards could be seen as a performance assessment tool for practitioners and that this could be problematic. The managers addressed this philosophically by sharing that the focus of the tools was on each child or youth and ensuring their progress, not on job performance. More practically, practice experts worked with managers to develop a set of resources specifically addressing each recommendation, so practitioners could easily access and review how to implement what was being suggested.

Another key challenge is *having enough data to scale this type of work*. Many nonprofit providers do not have the 250+ cases of longitudinal case-level data to get started. However, as the cost of program administration systems and the number of vendors grow, there are many more organizations that have and/or are in the process of setting up and implementing robust program administration data systems. So, there are many more organizations that are getting ready or already there. The learning networks described above also may provide on-ramps for organizations not yet ready to create their own modeling and tools.

A fourth challenge is *ensuring that all identifying data are protected and secure*. Technologies are now in place that, whether within organizations or in the cloud, protect the identity of cases in datasets. Organizations that could be a part of a learning network would not have to share data, but could keep it secure behind their own firewall or behind the secure firewalls of HIPAA-compliant cloud platforms like Amazon AWS. The learning hubs

could leverage aggregate findings. If there is a desire to share data, there are efforts of organizations like Brighthive[3] that create data trusts, which, both technologically and through policy agreements, create shared data architectures and processes that protect information.

PROVIDING INSIGHTS IN REAL TIME

Gemma Services has begun implementing their dashboards, and practitioners now are using actionable evidence in real time. The dashboards and tools resulting from precision modeling are designed to be actionable not just for practitioners but also for program developers and managers, organizational leaders, and system change agents, including policymakers and funders. Additionally, the goal of these tools is to improve program delivery in a tailored way that will improve outcomes. The evaluation tools will provide the insights, and the resource guides and the learning and professional support of peers and managers will help practitioners take the actions to increase the proportion of those who achieve the desired outcomes.

The precision analytics process and associated tools led to a culture shift in how practitioners at Gemma Services do their work and how performance is evaluated. Prior to the development of the learning and evaluation dashboards, progress and results of practitioner-led case-specific decisions were not available to practitioners (and were certainly not available in real time, updated on a daily basis). It was now possible to measurably view each child's or youth's progress, performance, and outcomes in a practitioner's caseload as well as the performance of their whole caseload, with updates on a daily basis.

Some of the early results of having immediate and up-to-date actionable evidence include:

- **Practitioners use case-specific dashboard evidence for case engagement.** For example, clinicians working with children in the residential milieu at Gemma Services are using the individual dashboard to understand more about the needs of the children they are working with on a daily basis.
- **Program directors use cluster evaluation information, including the names of children at different current outcome levels, for case planning.** This allows directors to meet with front-line staff to investigate why different youth are in different

places, examine what the data are beginning to indicate, and, most importantly, to more qualitatively examine the root causes for a child's outcome status. Program leaders find these inquiries, made possible by having up-to-date evidence, are strengthening case plans. Real-time feedback also allows for rapid plan changes.

- **Precision tools are motivating program directors and front-line practitioners to gather more data.** In the past, both organizations faced challenges with getting assessments completed on a regular basis. Now, with dashboards that update in real time, including indicating the date when the last assessment was conducted and/or data was entered, practitioners and program managers are visually cued to gather data more frequently.
- **The precision tools are engaging practitioners to help improve the data.** As more program directors, managers, and practitioners make deeper qualitative inquiries into their cases and what is going on, they are beginning to realize they need additional data points to test hypotheses that cannot be answered by the current data. For example, the Gemma Services practice expert learned that program directors felt strongly that parent engagement was a critically important variable that was not currently being captured in the data. They hypothesized that the quantity and quality of parent engagement, both with the child and the clinical staff, were key determinants of achieving many of the goals needed to reduce acuity. So, the practice expert worked with the program directors to design a set of questions that will be tested and modeled in the near future for use on an ongoing basis.
- **Precision modeling findings are being leveraged for policy change.** First Place for Youth's director of public policy, vice president of learning, evaluation, and strategic impact, and academic research partners have used the findings from their precision modeling process to write a policy brief, Raising the Bar: Building System and Provider-Level Evidence to Drive Equitable Education and Employment Outcomes for Youth in Extended Foster Care.[4] The purpose of this paper is to encourage the state of California and federal policymakers to scale First Place for Youth's extended foster care model, in conjunction with their Youth Roadmap Tool learning system.

REFLECTIONS

Gemma Services is raising funds to scale and create a learning network of similar organizations everywhere, all learning together by adopting and/or building their own learning systems. In fact, Scattergood Foundation plans to fund Gemma Services to scale their models to other residential mental health providers throughout the region, state, and country.

Gemma Services' vision is to *become a collective learning hub for similar programs* in communities throughout the United States. First Place for Youth is beginning to plug new affiliate programs throughout the United States into their Youth Roadmap Tool. If a similar program doesn't have enough data, they could, for example, begin by adopting First Place for Youth's extended foster care question sets, algorithms, and dashboards until they have enough longitudinal data to build their own context-specific precision models and tools. If they have enough of their own program administration data, they could build their own precision models. At this point, the build cost is affordable to most larger organizations, and the ongoing support and maintenance cost is sustainable for medium to large organizations.

It is important to note that both organizations voiced a goal of *eventually engaging beneficiaries* in their precision modeling process. However, they wanted to focus on the practitioner for their first precision projects to better understand what the process and engagement actually entailed, to better inform and plan for engaging beneficiaries and their families. Both organizations still are primarily focused on practitioner utilization. That said, Gemma Services has begun developing data gathering instruments to gather information from parents as to their engagement with their child's residential treatment.

NOTES

1. This step is analogous to propensity score matching (PSM) statistical procedures use in tens of thousands of health, education, political science, economic, etc. peer-reviewed observational studies to minimize selection bias. However, training machine learning algorithms to identify matched comparison groups mitigates a significant problem that leading social science researchers and statisticians Gary King from Harvard and Richard Nielsen from MIT, identified in their 2019 paper "Why Propensity Scores Should Not Be Used for Matching," *Political Analysis* 27, no. 4 (2019), pp. 435–54, proving that PSM creates experimental imbalance.

2. This machine learning process, using simple decision tree algorithms, adheres to King and Nielsen's recommendation of using fully blocked matching instead of PSM.

3. See Brighthive's details, https://brighthive.io/.

4. "Raising the Bar: Building System and Provider-Level Evidence to Drive Equitable Education and Employment Outcomes for Youth in Extended Foster Care," https://firstplaceforyouth.org/research-brief-raising-the-bar/.