Centre for Public Impact
A BCG FOUNDATION

CHILD PROTECTION

ARTIFICIAL INTELLIGENCE
in government

The Allegheny Family Screening Tool

A case study on the use of AI in government

In brief

- This is a case study that examines in detail a single application of data analytics in the public sector, in this instance an algorithmic decision support system for child protection in Allegheny County, Pennsylvania.
- The county's child protection agency was under fire in the 1990s after serious failures in its duty of care to local children. The incoming director embarked on a series of reforms, including extensive computerisation.
- The agency worked with a consortium of academic researchers to implement an algorithmic tool for call screeners. The screeners have to decide whether to investigate information received during calls from members of the public or professionals, who report their concerns about children who may be at risk.
- The tool uses data mining techniques to search for patterns in historical data, which then fed into the predictive risk model.
- Although the call screeners initially felt threatened by the tool's introduction, they came to find its risk predictions valuable in supporting their decision-making.

To learn more about how to proceed along a path towards using AI in the public sector effectively and legitimately, read our other AI case studies and our accompanying paper: How to make AI work in government and for people.

The Challenge

Allegheny County is located in the state of Pennsylvania and comprises the greater metropolitan area of the city of Pittsburgh. In 2016, the poverty rate in the county was 11.6 percent, just below the nationwide rate of 12.3 percent.[1][2]

Once known as "the steel city", Pittsburgh's steel mills have vanished from the economic landscape, and it has since been praised for its economic diversification with a focus on sectors such as services, technology and education. However, Allegheny County as a whole is also named among the top 10 least equal counties in Pennsylvania when it comes to income.[3]

In 1994, news items about a small toddler beaten to death by her father because she would not stop crying horrified Allegheny County's residents. The attention of the press and public quickly shifted to the county's child protection services (CPS), asking how it could have granted custody of the toddler to the abusive father, after she was previously in foster care. An official investigation concluded that child welfare or child protective services were partly responsible for the child's death, damaging the agency's already low public standing, labelling it "a national disgrace".[4] In an effort to restore trust via a first-ever town hall meeting in 1996, the new Allegheny County child welfare office director Marc Cherna and his staff endured parents calling them "child-snatchers" and accusing them of being the reason they could no longer see their children.[4][5]

Before 1996, apart from experiencing severe public mistrust, Allegheny's County's child welfare office was also lagging behind in its use of technology. Allegheny County used a few computers for administrative tasks only, and most records of children in foster care still existed only as paper copies. By 1997, a year after his start date and with the backing of other community leaders, Marc Cherna became director of a newly-formed Department of Human Services (DHS). He embarked on a series of reforms to integrate human services in Allegheny County in order to "lead it into the 21st century". By 1999, the Cherna team had established the Allegheny County Data Warehouse. It brings together and integrates client and service data from both internal and external sources. "It was created by consolidating publicly-funded human services data (e.g., behavioral health, child welfare, intellectual disability, homelessness and aging) and, over time, expanded to include data from other sources (e.g.: information about court cases and data from public schools)."[6] The Data Warehouse allowed child welfare caseworkers and staff in other DHS offices to see "if their clients lived in public housing, received unemployment benefits and food stamps, or had grown up in foster care, without spending time running down records from other offices"[4].

Marc Cherna also brought a culture of transparency and openness to DHS from the moment he started through a variety of mechanisms, including a regular research and publication agenda, and the 'Director's Action Line' which allows the public to register their concerns and questions about DHS services.[7][8]

Overall, through the provision of better data via the Data Warehouse, caseworkers helped to tackle the five-year backlog of child welfare cases, providing faster and more effective services. Additionally, Cherna encouraged evidence-based decision-making in order to intervene at the most effective moment. However, around 2014, he and his team decided to take an even bigger step – to create a more robust and data-informed system of screening reports of child abuse.[5]

In the US, child abuse and maltreatment investigations are generally triggered by referrals, by someone calling in to child welfare agencies or hotlines. Although referrals are typically made by medical staff from hospitals or teachers at schools, allegations can also be raised by anyone else, including noncustodial parents, landlords and neighbours.[9][10]

All 50 states have their own child abuse and neglect reporting laws that in some way mandate people to refer suspected maltreatment to a CPS agency, most typically via a call centre under the management of the respective CPS agency.

When a call comes in, screeners need to decide rapidly whether an investigation needs to begin, based on the information from the call they have received. "Call screeners are expected to render a decision on whether or not to open an investigation within an hour at most."[9] If they decide that an investigation is not warranted, the call will get "screened out", meaning no further investigation will take place (increasing the risk of an 'at-risk' child falling through the cracks of the system with no intervention).

Even experienced staff who are making decisions based purely on individual allegations rarely have the time to investigate and research a case history fully, and have to rely largely on the information received during the initial call.[10] This leaves screeners susceptible to making decisions that are swayed by their own opinion biases and beliefs. Many decisions on whether to screen an allegation in or out are also difficult to make because the individual allegation is hard to judge. And the decisions made can have a high price; for example, the number of children dying from neglect remains worryingly high. In 2016, 1,750 children are estimated to have died under such circumstances in the US as a whole,[11] an increase of 213 deaths compared to a recorded 1,537 child fatalities in 2010.[12]

The Initiative

In 2014, with the help of funding from a federal grant and local foundations, DHS sought proposals for "innovative solutions" to help "improve the delivery of services to its consumers by using data to improve decision making".[13]

DHS based its objectives in the Request for Proposals (RFP) on developing a tool that could be replicated, without much customisation, in other jurisdictions.[13] DHS deliberately worded the RFP so that it was open to new technologies and techniques that would help address the issue at hand. Specifically, DHS was looking to "explore its options".[10]

After receiving 17 applications, DHS selected a consortium of researchers from Auckland University of Technology (New Zealand) the University of Southern California, the University of California at Berkeley, and the University of Auckland (New Zealand). The consortium's proposal included a concrete use of Predictive Risk Modelling (PRM) to support decisions made at the time a child had been reported for alleged maltreatment.[14]

Erin Dalton, Deputy Director of DHS's Office of Data Analysis, Research and Evaluation, said of the successful proposal: "They really impressed us with having the values that we have... they're very concerned about ethics... as opposed to some of the other big boys around, who had their model and wanted to sell it to us."[10]

The successful consortium identified three broad objectives for the new tool:

1. Improve the ability to make efficient and consistent data-driven service decisions based on available records
2. Ensure public sector resources were being equitably directed to the county's most vulnerable clients
3. Promote improvements in the overall health, safety and wellbeing of county residents.[14]

The Allegheny Family Screening Tool (AFST) has been built by using data mining techniques to search for patterns in historical data. Its purpose is to support DHS call screeners to make data-informed decisions in response to calls alleging child maltreatment rather than the original decision-making system, which put a lot of value on the allegation made by an individual and the judgement of the screener.

The Tool generates a screening score for each child on a call and the score is displayed in a vertical, coloured bar, which colour codes numbers from 1 (low risk, in green) to 20 (high-risk, in red). The screening score indicates the likelihood of re-referral or out-of-home placement within two years. The higher the score, the greater the risk of an adverse event.

If the judgment of the human screener and the screening score diverge on the question of risk the screener has the option, together with the supervisor, to override the screening score.[10][15] This means, that call screening staff have discretion to decide whether the screening score matches the situation.[16] For example, some "children have a long history of juvenile probation, mental health services and parents with mental health services but there's no real need for [DHS involvement] and a lot of times those end up being mandatory scores so we do override those if there isn't a need for us to be involved".[16] This implies that the screening score is not making the decision for screeners but helps them to feel more confident. "It's not making the decision for us. Situations in which there's a gray area and you're certain on how to proceed, it makes you feel confident in your decision or guides you in a direction"[16].

> The Family Screening Score provides additional information for the call-screening decision-making process, so the tool is known as a decision support tool. It does not replace clinical judgment. The Family Screening Score is only intended to inform call-screening decisions and is not used to make investigative or other child welfare decisions.[17]

The score is known only to screeners and their supervisors and is not revealed to the families, investigators or judges. Maintaining careful control over the dissemination of screening scores was identified by ethicists as an essential step towards reducing stigma. "If they know this family is a 20, they are going to be much more likely to recommend removal of this child than if they knew the score was lower".[10]

Erin Dalton explained her reasoning behind being as open as possible: "I'm very against using government money for black-box solutions, where I can't tell my community what we're doing."[18]

DHS has complete ownership of the algorithm and kept it open-source to make it easily shareable with other jurisdictions, and also invited independent process and impact evaluations of the tool, which helped to strengthen legitimacy and transparency.[15][16][18] Furthermore, Marc Cherna, Erin Dalton and the research team engaged in community dialogue to inform the public and give them the chance to ask questions.[7][16][19]

A key reason for developing the AFST was to improve the use of information already available to call screeners to help them make better decisions. However, screeners had mixed feelings about the new tool when they started using it. They had limited involvement in the development process when researchers designed the PRM software, and developers were bound by the internal process of the county and their staff mechanisms.[15] In hindsight, all parties agreed that more internal staff involvement during the development phase would have helped those who use the tool on a day-to-day basis to understand how the tool works and what data it uses. An early engagement would have also benefited the developers, who heavily rely on staff feedback to make the tool more user-friendly.[16]

Some staff initially felt threatened by the tool, thinking that the new technology would ultimately make their work obsolete. Some also wished for more training explaining the tool, which would have helped them understand how to interpret the score. A frontline staffer currently working in call screening summarised this lack of engagement: "[We] did not really understand what it was and how it was going to impact [our] jobs. If it would have been better explained [at the start] people would have been more accepting and less afraid." [16] This sentiment was also shared by the lead researcher, Rhema Vaithianathan, who said: "I would say that deep engagement with the frontline staff possibly would have led to more effective use of the tool from the start".[15]

However, feedback from screeners was elicited after the initial development phase. "Everybody's opinions have been used to make the smoothest, most user-friendly process possible", and staff feedback also contributed to the tool's success.[16] In addition to internal responses, DHS placed great importance on community feedback. As they had done previously when establishing the Data Warehouse in the 1990s, DHS did a lot of work in the community to engage different groups – from lawyers, to former foster children and parents.

Because it is critical to minimize the adverse effects of identification of children as 'at risk', including possible stigmatization, the County commissioned, and responded to, a full independent ethics report prior to implementation.[9][10] The authors concluded that implementing the AFST was ethical, but also that not using it might be unethical, saying: "it is hard to conceive of an ethical argument against use of the most accurate predictive instrument".[9]

Furthermore, DHS recognized a general discriminatory bias in public databases, including its own Data Warehouse. Erin Dalton told the New York Times that: "All of the data on which the algorithm is based is biased. Black children are, relatively speaking, over-surveilled in our systems, and white children are under-surveilled. Who we investigate is not a function of who abuses.

It's a function of who gets reported".[9] On this background, many officials in Allegheny County agree that "by adding objective risk measures into the screening process, the screening tool is seen [...] as a way to limit the effects of bias".[9]

The Impact

The AFST has been in use since 2016. Preliminary data suggests that "average screen-in rates have been nearly identical to rates for the same period of time one year prior (approximately 43 percent)".[18] "The system, after several rounds of improvements, has achieved 77 percent accuracy in determining whether a child is at risk of placement, based on a set of previous cases from 2010-2015. This has increased the efficacy and efficiency of the call screening process: many children who would previously have been screened out, even though they had a high risk of being placed in foster care within two years, were now screened in, and vice versa."[21]

Furthermore, the proportion of low-risk cases being screened in for investigation dropped from half to one-third, meaning that high-risk cases were being investigated ahead of those with a lower risk.[9] Marc Cherna himself sums it up:

> *"I don't think we're necessarily saving money, but we're using our resources better, in that workers are going out on higher risk cases and not going out on lower risk cases because we have limited resources. The demand always outweighs the supply."*[10]

This confirms his aim stated early on of using "data to support families earlier, so that child welfare does not have to intervene".[10]

While call screeners were initially afraid that a predictive risk modelling tool for screening threatened their jobs, they learned to feel very differently about it. "Having access to all that information is very beneficial, and using everything in your power to make the most informed decision and predictive analytics is part of that. I can definitely see the benefit."[16] Staff felt empowered when they understood that the score was supporting their decision-making, rather than making the decision for them. The tool "makes you feel justified in your decision-making... There are situations in which the score does not always match the situation, and those times we are able to override, so it's not making the decision for us."[16]

The AFST has mainly been criticised for using biased data. "There is widespread agreement that much of the underlying data reflects ingrained biases against African-Americans and others",[9] because there exists a tendency to over-report and collect data on children from ethnic minorities. For example, "in 2015, black children accounted for 38 percent of all calls to Allegheny County's maltreatment hotline, double the rate that would be expected based on their population [and actual abuse rates]".[9] Furthermore, the Data Warehouse has access only to publicly-funded insurance (e.g. Medicaid) data and does not have information about private insurance, meaning, for example, that information about a person using a public insurance like Medicaid to get treatment for illegal substance abuse is captured in the data, while information about a person with private insurance getting the same treatment is not.[10]

However, 16 months after the tool went live, it was reported that families were being treated in a more consistent manner than before. The research team also found that the effect of insurance type on data availability was not as large a problem as previously thought. One of DHS's key learnings is that it is possible to achieve the same results by feeding fewer data fields into the algorithm. This potentially questions the need for a sophisticated detailed personalized Data Warehouse versus more general universal data to achieve results with the same accuracy. "What we're finding is you don't need all of this data. You can drop key variables and still have an accurate model. We're still working on that but that's something we absolutely did not know".[10] This suggests that future applications of predictive risk modelling may need less data but as much universal data as possible, to achieve sufficient accuracy while decreasing the risk of bias.

On paper, at least, the bias against African-American children has decreased, and "the machine learning world is getting much more sophisticated about how to detect discrimination and bias in these algorithms".[10] Because the tool is acknowledged to be using biased datasets, Marc Cherna and his team wanted to explain their reasoning to the public. They placed a very high importance on community engagement, transparency of algorithms, and independent quality control. They attended town hall meetings and gave interviews, which appeared to pay off with greater support from the public. As Cherna explained, "we did a lot of work in the community for a couple of years".[10] He and his team talked to people from across the community: the parents' lawyers, the children's lawyers, the parents themselves, and people in the system or previously in the system. "And they understand it and they feel if we can improve our services, we should."[10]

However, concerns still persist. Some parents have expressed worries about how the tool could impact them if they get ranked as a "false-positive" – "the model rating their child at high risk of abuse or neglect when little risk actually exists".[22] Furthermore, negative reports and studies on the bias in the data caused one author to say that the tool "views parents who reach out to public programmes as risks to their children",[23] because that will become a datapoint, which will ultimately impact the scoring. While one of the lessons from Marc Cherna and his team is that using fewer data that are more universal does not significantly reduce accuracy of the scoring, it seems that at least some parts of the public remain unconvinced.

This case study is part of our series of AI use cases in government. It illustrates how the public sector can employ AI in social welfare provision. To learn more about how to proceed along a path towards using AI in the public sector effectively and legitimately, read our accompanying paper: How to make AI work in government and for people.

BIBLIOGRAPHY

[1] QuickFacts - Allegheny County, Pennsylvania, United States Census Bureau, 2017

[2] Poverty 2017, United States Census Bureau, 2017

[3] Income Inequality in Pennsylvania Skyrocketing, Keystone Research Center, 2018

[4] Laura Santhanam, Can big data save these children? PBS, 2016

[5] Sam Gill, Indi Dutta-Gupta, and Brendan Roach, Allegheny County, Pennsylvania: Department of Human Services Data Warehouse, Data-Smart City Solutions, 2014

[6] Allegheny County Data Warehouse, Allegheny County, 2018

[7] OF AND FOR THE COMMUNITY: How Community Engagement and Communication Helped to Shape the Allegheny County Department of Human Services, Allegheny County, 2016

[8] From Almshouses to Excellence: A History of Child Welfare in Allegheny County, Allegheny County, 2015

[9] Dan Hurley, Can an Algorithm Tell When Kids Are in Danger? New York Times, 2018

[10] Interview with Marc Cherna and Erin Dalton, 2018

[11] Child Maltreatment 2016, U.S. Department of Health & Human Services Administration for Children and Families Administration on Children, Youth and Families Children's Bureau, 2016

[12] Child Maltreatment 2010, U.S. Department of Health & Human Services Administration for Children and Families Administration on Children, Youth and Families Children's Bureau, 2010

[13] Request for Proposal to design and implement decision tools and predictive analytics in Human Services, Allegheny County - Department of Human Services, 2014

[14] Rhema Vaithianathan, Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation, Center for Social data Analytics, 2017

[15] Interview with Rhema Vaithianathan, 2018

[16] Interview with frontline screeners, Allegheny County, 2018

[17] Predictive Risk Modeling in Child Welfare in Allegheny County The Allegheny Family Screening Tool, Allegheny County, 2016

[18] Rachel Courland, Bias detectives: the researchers striving to make algorithms fair, Nature – International Journal of Science, 2018

[19] Child Abuse Statistic in the U.S., American Society for Positive Care of Children, 2015

[20] Allegheny Family Screening Tool: Frequently Asked Questions, Allegheny County Department of Human Services, 2017

[21] Kendra Schneider, Data for Social Good Case Study: Allegheny Family Screening Tool, SIX, 2018

[22] Virginia Eubanks, A Child Abuse Prediction Model Fails Poor Families, Wired Magazine, 2018

[23] Richard Wexler, Many problems with the Allegheny Family Screening Tool, Pittsburgh Post Gazette, 2018

Centre for
Public Impact
A BCG FOUNDATION

Join the conversation

ai@centreforpublicimpact.org
Follow us @CPI_foundation: #FindingLegitimacy
#AILegitimacy

October 2018