Ethical Evaluation

of the

Predict-Align-Prevent Program

Professor Tim Dare
Philosophy
University of Auckland

December, 2018

THE UNIVERSITY OF
AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Executive Summary

This report sets out what I take to be the principal ethical issues raised by the Predict-Align-Prevent Program (PAP), and to make recommendations which I believe would help to address or mitigate those issues. [1]

I am satisfied that the PAP program has the potential to deliver genuine benefits while avoiding some of the familiar risks of alternative approaches to targeting child protection services.

While I identify a number of grounds for ethical concern, I am satisfied that, on balance, the benefits of the PAP program would outweigh the risks it poses. It is my view that implementation of the Predict-Align-Prevent Program can be justified from an ethical perspective.

Introduction: Report Purpose and the Predict-Align-Prevent Program

This report offers an ethical analysis of the Predict-Align-Prevent Program. I do not comment on the technical detail of the geospatial predictive risk model, which is central to that program.

1. As its name suggests, the Predict-Align-Prevent Program has three phases:

   1.1. **Predict:** During a 'predict' phase, PAP uses geospatial predictive risk modeling to identify high-risk geographical locations[2] based on environmental features. Geospatial modeling relies upon the idea that spatial externalities can be used to predict the locations of future events and that those externalities can be ranked according to their contribution to the probability of the predicted outcome.

   1.2. **Align:** During an 'align' phase, PAP aims to use the predictive information about the relative locations of future child maltreatment events and proximate risk and protective factors to identify opportunities to work strategically with communities and providers to align services, education, and resources to locations where they are most likely to reach children at risk. In the Align phase, PAP utilizes the predictive maps overlaid with public health and community asset locations. Existing prevention service delivery can be overlaid on high risk areas to evaluate spatial allocation and capacity needs. During the Align phase, the leadership and project managers work with existing community leaders, stakeholders, and coalitions to align and augment existing prevention efforts. The aim is to ensure that "the current supply of child welfare services is properly distributed relative to the demand for these services."

[1] This ethics analysis was made possible by support from Casey Family Programs.

[2] The model supplied by way of illustration uses a grid with 1000 x 1000 ft^2 cells.

1.3. **Prevent:** During a ‘prevent’ phase, PAP aims to generate baseline data and to actively surveil risk, protective, and outcomes data in high-risk areas to measure the effectiveness of particular implementations of prevention programs in those areas, and to inform future prevention efforts.

2. So described, PAP differs from many social policy uses of predictive analytics in that it is *place*- rather than *individual- or family-based.* Although the program appears to leave open the *nature* of the service, education, or prevention programs engaged with or put in place during the Align and Prevent phases,[3] the overall focus of the program does not require the use of what is in some jurisdictions called ‘personal information’.[4]

   If an implementation decision consists, for instance, in increasing the concentration of protective factors in areas identified as high-risk, there may be no need to explicitly engage with or know the identity of at-risk individuals or families. If what I take to be a premise of the Predict phase of PAP is correct, namely that there is a correlation between factors such as the concentration and proximity of risk and protective factors and the risk of maltreatment, then one might alter maltreatment risk by decreasing factors shown to be correlated with risk and increasing factors shown to be protective in an area. Doing so would not require access to information about individuals or families who might benefit from those changes.

[3] See *The importance of implementation decisions*, paragraph 5, below.

[4] Where ‘personal information” is any information that could be used to identify the person the information concerns. There are obvious types of personal information – names, addresses, telephone numbers, social security numbers, for instance – but information can become personal information if it can be linked to an individual when combined with other available information: a description of a person as someone who had a particular unusual medical procedure last week could be personal information if someone could use that information, together with information about which hospitals offer that procedure, and the schedule of procedures performed at those hospitals, to work out who the person is.

Ethical Evaluation.

3. Data integrity

Data integrity is an issue for all data uses, and especially important where data analytics might drive decisions which have the potential to influence significant decisions about individuals, families, or communities.

There appears to be at least one obvious ground for concern about the integrity of the data underpinning the Predict phase of PAP. The model uses existing reports of child maltreatment to produce its original maps. The integrity of those maps, then, depends upon:

a. the substantiation of the original cases, and

b. confidence that reports of child maltreatment events are not skewed by, for instance, bias or uneven surveillance.

PAP should, I think, be aware of these threats to the integrity of the data upon which the initial maps rely and have an account of steps which have been and will be taken to ensure that their influence is avoided or mitigated.

Future data integrity may also be an issue if, for instance, PAP itself leads to increased surveillance or reports in areas identified as high-risk. The issues are straightforward and acknowledged in the use of place-based modelling in law enforcement,[5] where identification of areas as high-risk led to extra police presence in those areas, resulting in increased detection of criminal activity, and – in addition to the imposition of the burdens of law enforcement on targeted populations – to problematic feedback loops which increasingly threatened to distance predictions from actual patterns of offending.

Whether the Align and Prevent phases of PAP might have those effects depends upon just what is involved in those phases. The issues are presumably not as straightforward in the child maltreatment case as they are in predictive policing, since the services, education, resources, and prevention and protection programs which are the focus at the Align and Predict phases of PAP are likely to be less directed to surveillance and detection than those deployed in law enforcement.

Nonetheless, it is important that PAP consider the possibility that the increased allocation of child protection resources to high risk areas could lead to increased surveillance and reports in those areas, threatening the integrity of data which might be relied on in future modeling.

4. Increased surveillance as potentially intrinsically wrong.

I have mentioned the threat increased surveillance poses to the integrity of the data which might be relied on in future modelling. It is important to appreciate, however, that increased surveillance is in itself a burden to communities, independently of its tendency to taint data sets. Disproportionate surveillance may, therefore, be *intrinsically*

[5] The best known example is PredPol. There is a considerable literature discussing 'the 'feedback problem'. See, for example, see Kristian Lum and William Isaac, 'To predict and serve?, *Significance* 13.5 (2016): 14-19. https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2016.00960.x

and not merely *instrumentally* wrong.[6] It may be wrong even if it causes no further dis-value by, for instance, tainting datasets.

It is important that PAP consider the possibility that the increased allocation of child protection resources to high risk areas could lead to increased surveillance in those areas.

5. The importance of implementation decisions.

Whether the use of predictive analytics in social policy contexts is ethical always depends to a significant degree upon implementation decisions. Decisions about how predictions are *used* in engagement with individuals, families, or communities will typically determine whether the use of predictive analytics as part of a policy process is ethical or not.

To its credit, PAP includes commitment to significant implementation components: The Align and Prevent phases are implementation phases.

However, just what those implementation components will involve in the PAP program is at this stage left unspecified.

This lack of detail is perhaps unavoidable. PAP intends to *identify opportunities* to work strategically with communities and providers during the Align phase, and to measure the effectiveness of particular implementations and inform future prevention efforts during the Prevent phase. Given these goals, it may be impossible to say very much in advance about just what future implementations will look like, and so difficult to say very much about how effective or (for instance) intrusive they are likely to be.

The point for the moment is not that one cannot offer an ethical evaluation without a complete account of the likely implementation decisions which may be made at the Align and Prevent phases of the program. Instead, it is to point out that those answers, when they are available, are important to an ultimate ethical evaluation.

I give some examples of the way in which implementation decisions may bear upon the ethical evaluation below.

For the moment, I suggest that PAP monitor implementation decisions made during the Align and Prevent phases, and appreciate their potential ethical implications in light of the more specific comments made elsewhere in this evaluation.

6. Privacy

As noted, place-based predictive modeling appears to have some obvious ‘privacy’ advantages over individual or family-based modeling, since it does not rely upon detailed information about individuals or families: “identifying details about vulnerable and maltreated children …, such as name and social security number are not necessary …”.

However, there are caveats and concerns around these ‘privacy aspects’ of the PAP program. Some of these are related to the robustness of the protection of the anonymity of individuals, families, and households.

[6] Something is *intrinsically* valuable (or wrongful) if it is valuable (or wrongful) independently of any further value (or disvalue) it might produce. Something is *instrumentally* valuable (or wrongful) if its value (or wrongfulness) depends upon its contribution to other valuable (or wrongful) things or states of affairs.

- Some of the input data, such as records of maltreatment events, will include address level data. That address level data is to be geo-coded and placed within maps of the center of interest. The resulting maps will place events within 1000sqft units and will not include identification of individual addresses.

  It is important that the process does include significant barriers to the recovery of address level data, and that there is a process for managing any privacy breaches which do occur.

- It seems quite likely that individuals, families, and households will occasionally be identifiable (even if not identified) by the combination of geospatial data available to those who have access to the predictive model, and other information which may or may not be controlled by PAP or other stakeholders; it seems likely, that is, that it will occasionally be possible to make de-identified data into personal information by combining it with other information available to those involved in PAP or other stakeholders.

  Again, it seems important that PAP have a process for making such re-identification difficult, and for managing any privacy breaches which do occur.

- I note that the process does include identifying and removing duplicate events. If that process does not include address level data, it seems likely to suffer from at least some degree of imprecision. More than one maltreatment event may occur within a 1000sqft area, and perhaps (if more than one child is involved) at more or less the same time. Presumably such events would be rare, and perhaps do not pose any significant threat to the accuracy of the model.

However, I am not inclined to make too much of these issues around the robustness of the protection of privacy under PAP. Even allowing for the possibility that it will occasionally be possible to identify individuals, families, or households, the risks seem fairly low and manageable, especially when compared to those which are likely to occur under plausible alternatives to PAP's place-based modelling.

Other caveats and concerns around these 'privacy aspects' of the PAP program focus less on breaches or failures and more on structural features of the PAP program.

- The Predict phase of PAP does not require access to personal information. Some implementation decisions may, in some sense, also be place-based. If, for instance, implementation were to involve no more than changes to the distribution of risk and protective factors, there may not need to be any contact with identifiable individuals: PAP might be able to eschew 'personal information' at every phase. Some descriptions of the Align phase suggests that this is what the program has in mind. The case study supplied to illustrate the model says:

  > "Align: Using the knowledge from the predictive model, we develop a strategic planning framework, embedding the risk predictions into a larger analysis describing whether the supply of child welfare services is properly aligned with the demand for these services."

However, other descriptions of the Align phase suggest that there may be cooperation with independent services and stakeholders; and so it seems possible that some of the services and preventive efforts encountered in the Align and Prevent phases will involve contact with individuals and families: they might require access to personal information. If that is right, at least some of the advantages generated by placed-based prediction may be offset by targeted interventions at later phases of the program.

Again, I think this requires careful attention to particular services and implementation decisions made during the Align and Prevent phase to ensure that they do not undermine the ethical advantages of the place-based character of the PAP program.

7. Stigmatization

The identification of areas as high-risk for child maltreatment may associate those areas and their occupants with conduct widely regarded as shameful and improper. Hence, there seem to be grounds for concern that the PAP program may lead to the ***stigmatization*** of geographical areas and those who live in them, to the marking of those areas and the people who live within them as flawed and less worthy of respect than others.

Two obvious reasons to think such stigmatization improper are:

a) That the process is *predicting* improper behavior. In the case of predictive models, stigmatization turns on behavior that has not occurred. Any burden of stigmatization in these cases is allocated in advance of confirmed wrong-doing.

b) That most of the residents of areas stigmatized will be – and remain – entirely blameless. They – adults and children alike – will bear the burden of stigmatization nonetheless.

Less obviously, it is *almost* always[7] improper to place the burden of stigmatization *even upon those who are not blameless*. Not only does stigmatization go beyond simply assigning judgements of guilt for specific behavior, marking individuals and communities as less worthy of respect, it is likely to be unproductive. It may be difficult to establish the sorts of relationships necessary to implement effective protective engagement with individuals and communities who feel cast out and see that they and their community are judged to be flawed. Stigmatization is likely to be a barrier to engagement.

The threat of stigmatization is not unique to place-based predictive modeling, or indeed to predictive modelling in general. It might seem, indeed, that place-based modelling has some advantages over individual or family-based modeling with regard to stigmatization, since it does not identify individuals or families. On the other hand, place-based modeling potentially stigmatizes entire areas and all of the individuals and families who live within them.

[7] There is some interesting literature on when it might be legitimate to use stigmatization as a public health tool (e.g., S. Bayer. Stigma and the Ethics of Public Health: Not Can We, but Should We. Soc Sci Med 2008; 67: 470.; Andrew Courtwright , Stigmatization and Public Health Ethics Bioethics, 27.2 2013 pp 74–80) but it does not apply to the current case.

Stigmatization often works by reinforcing existing prejudice and stereotypes. Those who are already marginalized, who bear the burden of social disapproval – perhaps because they fall into stigmatized social, ethnic, or economic groups – are most vulnerable to additional stigmatization. Others are primed to think ill of them and likely to interpret or misinterpret information, such as an assessment that a neighborhood is judged to be high-risk for maltreatment events, in ways which confirm antecedent prejudices.

To the extent that the areas which are identified as high-risk at the Predict phase of PAP are subject to existing stereotypes and prejudice, the program has a responsibility to guard against stigmatization at the outset, and to identify ongoing mitigation for any stigmatization burdens the program does create or reinforce.

Stigmatization depends upon the communication of stigmatizing messages. Consequently, one way to guard against stigmatization is to control access to stigmatizing or stigma-reinforcing information.

That may be relatively straightforward with respect to some of the information produced by the PAP program. However it is possible that *many* of the initiatives undertaken at the Align and Protect phases of the PAP program will, in addition to whatever else they achieve, communicate judgements about the areas in which they take place. Focusing child protection resources in particular areas, for instance, may make those resources more available to residents but it may also signal to residents and outsiders that the area is 'marked' as an area in which children need special protection.

Here, as elsewhere, it needs to be acknowledged that these potential harms are not unique to the PAP program or to the use of place-based predictive modeling. Indeed, in the case study sent as background information, the *current* distribution of CPS offices across a city may already send the sorts of signals contemplated in the previous paragraph, and other approaches to delivering child protection services also threaten stigmatization.

8. **Pushing marginalized families beyond the reach of services**.

One familiar and specific worry about child protection programs which rely on identifying and targeting individuals and families is the danger that identification may *increase* risk to vulnerable children, since families, who will often already be marginalized – in part perhaps because of the sort of stigmatization discussed in the previous section - may choose to remove themselves from perceived or actual surveillance, hampering rather than improving delivery of services. Place-based modelling may reduce these risks, since families and individuals are likely to feel less specifically targeted in prediction or service delivery phases of the program.

Nonetheless, the PAP program should consider whether and to what extent the program might lead families to move beyond the reach of services and how to minimize that danger.

9. **Indirect Discrimination**

Discrimination may be *direct* (a policy or practice might, for instance, explicitly exclude a group from some service), or *indirect* (a policy or practice might, for instance, include eligibility criteria which are extremely unlikely to be met by members of a group, and so exclude members of that group from the service even though it does not explicitly mention them). For example, an employment policy which requires applicants to be over 5ft 11in tall might discriminate against women, indirectly, since men are much more likely than women to be over 5ft 11in.

Indirect discrimination may be entirely unintended. An example which occurred to me while reading the example provided was the presence of bus-stops. I often visit a particular US city and use the city buses. I am almost always struck that bus patrons are not a very diverse group, and that I stand out: drivers and passengers occasionally look surprised when I board. My sense is that bus routes, at least in that city, serve mainly poor areas, perhaps where residents are less likely to have cars or perhaps to be provided with parking facilities as part of their work.[8] If that is right, then when the model uses the ostensibly innocuous concentration of bus stops as a variable, it may be tracking poverty and the groups who seem disproportionately to use buses.

There may be grounds to worry that the PAP program discriminates indirectly in this way, against groups who are, contingently, more likely to live in the areas identified as high-risk. With respect to the example supplied by way of illustration, it is pointed out that though “no variables directly measuring race or income are included in the models”, “[i]ncome and race are inextricably linked to many of the census and exposure features used in the model.” The explanation goes on to show that though “the model generalizes well to places with varying incomes but generalizes less well to neighborhoods of different racial composition.”

***PAP should consider carrying out a population analysis of areas identified as high risk to determine whether particular groups are disproportionately represented in those areas. This will help obtain some sense of the risk that the program will discriminate indirectly against those groups*.**

10. Transparency.

For the most part, the model is strikingly and admirably transparent. It is proposed to develop a comprehensive open source framework for developing child maltreatment predictive models and to document a strategic planning process for converting maltreatment risk predictions into actionable intelligence that stakeholders will be able to use to allocate limited child welfare resources. The program documentation contains information on variables and methods. The program aims to be unusually transparent to child protection professionals and modelers.

It is widely accepted, however, that those who may be affected by the use of data relating to them are owed an account of how the data will be used. It is not clear that there are plans to make the program transparent to the residents of areas which are the subject to the modelling, and if so how the program will be presented and explained.

This concern may seem most pressing when it is proposed to use data *about* an individual or family, and where proposed data use will directly affect ‘data-subjects’. At the Predict phase, at least, the PAP program does not intend to use identified data about individuals (though, again, the program will use address level data about child maltreatment events which will be de-identified in the project and outputs), and under

[8] I also suspect that bus-patronage is at least to some degree stigmatized in that city.

at least some versions (those in which the interventions consist solely in changes to the concentration of risk and protective factors) the Align and Prevent phases will not involve individual or family level data either.

However, it seems very likely that people would be interested in knowing that geospatial modelling is being used to identify high-risk areas in their communities. They may well feel they have been shown a lack of respect if they are not informed about the program and how it works. There are likely to be some challenges around doing this effectively and in ways which do not alienate.

The PAP program should explore ways to increase transparency about the PAP program at least in those communities in which it is used.

11. **Consent**.

I note consent largely for completeness. Given that the only individual level data which will be used are the original and publicly available maltreatment records – data which will be anonymized before the modelling process – there is no obvious requirement for consent.

As elsewhere, I add the caveat that ***any services or protective programmes offered under the Align or Prevent phases of the program which do engage with individuals (e.g., which go beyond changes to the concentration of risk and protective factors in an area) may generate their own consent requirements.***

12. **Control and transparency**

The ambition to create an open source framework is admirable. It does, however, seem to create some risks both for PAP (and sponsors), and communities. Has consideration been given to the possibility that the framework might be misused, whether in an altered form, or by organizations or groups with less desirable policy goals than the PAP program and supporting organizations who may simply press for misinterpretations and misuse of the mapping capacity of the geospatial profiling components of PAP?

It may be that there is no effective way to prevent such misuse. Additionally, the benefits of an open source framework may outweigh possible harms, particularly once those harms are discounted by the probability of them actually occurring.

Nonetheless, it seems prudent for the PAP program to consider whether there are credible concerns about the possible misuse or misrepresentation of the program and how they might be managed.

13. **Need for education of those involved in implementing the PAP program.**

Leading on from the previous point, in the right hands the PAP program has the potential to do considerable good. However, it is easy to imagine information produced at the Predict phase (especially), and perhaps to a lesser degree, the services and programs engaged in the Align and Prevent phases being misused and misrepresented.

The PAP program team may be almost uniquely well-placed to explain the program – the significance and limitations of the information it produces at the Predict phase, the importance of the management of the interventions at the Align and Predict phases – to those who might take it up.

In my view there is an obligation upon the PAP program to ensure that those involved in implementing the program are appropriately informed about it. That obligation may be especially challenging given the laudable commitment to producing an open source framework. ***The PAP program should develop plans to ensure that they can meet these obligations, albeit with the understanding that some of the risks around this point may be outweighed by the importance of making a valuable tool as widely available as possible.***

14. **False positives, negatives, and other errors.**

Any sophisticated tool using predictive analytics is likely to be more accurate than approaches relying upon even guided expert judgement. However all predictive models will make errors: There will always be false positives and negatives; any model which categorizes cases into risk cohorts or bands will make errors around band margins; and there will always be a range of risk within risk bands.

These concerns are often extremely important where analytics may be used to drive decisions which have significance for individuals, families, or communities. Some of the normal concerns around false positive, false negative, and classification errors seem not to apply very clearly in the current case, since – depending upon implementation decisions – there may not be direct contact between the families and child protection workers.

The risks associated with mis-classification, whether because of false positives or negatives or because of errors within or around the margins of risk bands, will certainly be *different* from those associated with predictions focused on individuals or families. In the latter case, a particular individual or family will bear whatever cost comes with being labelled high risk, whether that is the stigmatization associated with being classified as at high risk for an adverse outcome, increased surveillance, or the burdens of intervention.

It is likely to be easier for individuals and families to avoid association with the negative consequences of mis-classifications when predictions focus on geographical areas, rather than individuals who will probably have to be identified to at least some people within a child protection agency.[9]

I assume errors are likely to affect classifications of risk and protective variables, the significance of proximity, and the like. Those errors may not be trivial, and of course should be guarded against, but they seem unlikely to result in clear cases of, for instance, wrong doing with respect to individuals or families, or communities, and to the extent that they weaken recommendations about the alignment or provision of protective services, they seem unlikely to have deeply significant consequences: one might contrast the threats posed by the use of place-based modelling in law enforcement which led to one on one encounters with law enforcement personnel, searches, arrests, and the like.

[9] This point will depend upon it being the case that the geographical areas are not reducible to individuals or families. As noted, the model supplied by way of illustration for the PAP program uses grids with 1000 x 1000 ft^2 cells. The ethical advantages associated with geospatial or place-based modeling over individual or family-based alternatives depend upon the unit of modeling not being reducible to households.

***My sense, then, is that while it is important that the PAP program take appropriate steps to ensure the accuracy of the modeling and appreciate that, at some level, error may begin to impact on the effectiveness of activities at the Align and Prevent phases, the PAP program is less threatened by inevitable false positives, false negatives, and other errors than most social policy uses of predictive risk modelling*.**

15. The use of Euclidean Distance.

The model uses Euclidean Distance to measure exposure to risk and protective factors. I wondered, however, whether distance may not always be the most relevant factor. Mightn't something like *accessibility* or *geo-temporal* proximity (i.e., how quickly and easily a person can get from one point to the other, which might be affected by things other than mere distance because of geographical features or transport systems) be equally or more important?

A risk-factor (a liquor store perhaps) might be easily *accessible* from a point because of a regular bus system even though comparatively geographically remote from that point, or comparatively difficult to access from a point because of some geographical feature (a river or a highway) even though comparatively geographically proximate.

Summary and Conclusions

16. Harms and benefits.

A key issue in assessing whether social policy uses of predictive modelling tools are ethical is determining whether, on balance, the benefits they deliver outweigh the adverse consequences they threaten or deliver. On the benefit side of the equation, I accept that the benefits which might be delivered by an effective child maltreatment prevention program are enormously significant – they warrant taking some risk and even imposing some certain costs – and I accept that the PAP program has the potential to contribute toward the development of effective child maltreatment prevention programs. I do note that little analysis has been done, or at least was provided to me, in support of these benefits.

I note that the PAP program does aim to monitor outcomes data in high-risk areas to measure the effectiveness of particular implementations of prevention programs in those areas, and to inform future prevention efforts. I think that it is important that it do so.

The PAP program does, in my view, threaten some adverse consequences, including, for instance:

- the potential stigmatization of neighborhoods;
- the possibility of increased surveillance (and hence the generation of reinforcing data);
- the possibility that marginalized families will respond to the program by removing their children beyond the scope of child protection services in that community, for example, by moving to another neighborhood;
- Potential indirect discrimination against groups who seem almost certain to be overrepresented in the neighborhoods identified as high-risk; and
- Some relatively minor threats to the privacy of information gathered at the Align and Prevent phases are not managed appropriately;

Some of these risks are relatively minor. Others can, in my view, be mitigated by way of the recommendations included in this report. Overall, for reasons discussed under the various sections above, I am satisfied that the benefits of the PAP program outweigh the costs it is likely to impose.

17. The Comparative Nature of Ethical Evaluation of Social Policy Uses of Predictive Analytics.

The previous point discussed the benefits and harms posed by the PAP program. But there is a related point: as is almost always the case with social policy uses of predictive analytics, many of the central questions around the ethics of the PAP program are essentially comparative, i.e., *about how the program compares, ethically, with alternative approaches*. I assume here that doing nothing about child maltreatment is not a plausible option and that at least some degree of targeting is both necessary, given resource constraints, and desirable, given the burden of some child protection initiatives.

If we are committed to at least some degree of targeting, there is reason to consider at least some forms of data analytics, and PAP's geospatial modelling seems in many respects an attractive alternative. Its focus on 'places' rather than individuals avoids many of the privacy and stigmatizing risks of individual or family based modeling, and I think reduces some of the dangers of the inevitable errors to which even the best predictive models are vulnerable.