Should algorithms be regulated like pharmaceuticals?

Learning lessons from the 2020 public examination results fiasco in the UK

Large-scale use of algorithms has the potential to cause severe harm to individuals and communities. Their use needs to be regulated effectively – perhaps learning from the Pharmaceuticals industry would be helpful.

> Learn to do good; seek justice, correct oppression; bring justice to the fatherless, plead the widow's cause.
>
> Isaiah 1:17 (ESV)

Introduction

As I write on 21st August 2020, we are still trying to grasp the full impact of the UK-wide chaos caused by algorithmic award of public examination grades following the cancellation of public exams during the Covid-19 lockdown.

The failure of the public examinations grade awards across the UK in 2020 offers a compelling case-study in the risks of large-scale public application of algorithms to decision-making (or recommendation). I believe this demonstrates clearly the need for robust, independent review and certification of algorithms prior to live use – *with focus on the impact of recommendations/decisions on individuals who are the subjects of these tools*. I think we could learn a great deal from the experience of the pharmaceutical regulation. We should approach licensing of large-scale algorithmic decision-making in a similar way to the approval of a new drug or medical treatment. There needs to be clear evidence that the defined approval protocol has been followed, evidenced and signed off prior to general use. That licence should state clearly the approved application for this set of algorithms.

This paper will not examine many of the details of the process followed to award grades – but it will discuss one or two key problem areas. It is not intended to add to the copious volume of criticism aimed at regulators, civil servants, members of the government or teachers. This is a cautionary tale that demonstrates we don't understand how to handle mass application of algorithms to sensitive areas of life; even though these kinds of approaches have been used for decades. I hope we can learn from this.

What happened?

Let's recap. In March, at the time of lockdown in the UK, the responsible Governments in England, Wales, Scotland and Northern Ireland cancelled this year's public examinations. Results would be awarded based on predicted outcomes. 'Results' had to be delivered from August 4th in Scotland and from August 13th in England. The various bodies had about 5 months to work out how this challenging task would be achieved. In 2020 there were about 720,000 A-level entries, and about 5.25m GCSE entries to be processed.

I have summarised the Exam Board objectives (slightly crudely) below:

1. maintain the 'value' of the exams; achieving acceptable comparability to previous years' results. (Avoiding massive grade inflation – nationally, regionally and for specific schools and colleges.)
2. handle large volumes of data, quickly and effectively – ensuring adequate standardisation across participating institutions.
3. providing fair outcomes to individual candidates, avoiding discrimination or bias.
4. remaining within the current legal constraints of algorithmic data usage –primarily expressed in the General Data Protection Regulation (GDPR).

Getting any one of these 'wrong' was inevitably going to cause a furore. However, any perception of unfairness was going to be particularly toxic. Results had demonstrably to be based on the available evidence gathered prior to lockdown including mock exams and show objectivity, transparency and a credible degree of predictability.

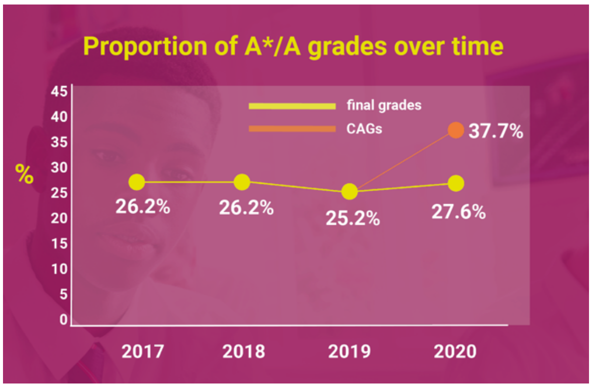
When the Scottish results were released on August 4th, the press quickly highlighted stories of seemingly strong candidates missing out on dream careers in medicine or other high-demand courses because of unexpectedly poor results. One 'unfair result' is one too many; but the evidence suggested there were thousands. The human consequences could have been devastating.

Unfortunately, politicians tend to get trapped by big picture concerns and forget the individuals who are the subjects of these processes. As a result, the English government seemed most concerned that they could relate this years' outcomes to preceding years. They assumed that the [relatively few] anomalies could be dealt with through (a) appeals and (b) in person exams sat this autumn or resits next year. Prior to announcing the results no one in Ofqual or the Department for Education seemed to recognise that their chosen approach could not deliver the level of individual accuracy required (though Jon Coles, the Chief Executive of United Learning clearly did – see below). I suggest that the instigators and operators of algorithms are rarely capable of marking their own homework. As algorithms become more complex and are used more widely, it is critical that we establish an appropriate but effective regulatory environment to oversee this: one that puts the needs of the individual first.

In Scotland, the decision to revert to Teacher Assessed Grades was taken just before the A-Level results day in England. Ofqual might have anticipated that they would be about to embark on an identical journey to the SQA – and that is just what happened.

Were algorithms necessary?

It is worth making sure we are clear why some kind of algorithmic adjustment was necessary. This part of an Ofqual infographic makes the case: [1]



In an ideal world, Centre Assessed Grades (CAGs) would have needed only a little moderation to provide a reasonable outcome as teachers know their own students better than anyone else. Reality is, that the unmoderated raw CAGs would have resulted in nearly 50% more A*/A grades than last year. These figures are implausibly high and represented a massive dilution of the perceived quality of 2020 A-Levels over previous years. This would have been politically unpalatable. Substantial statistical moderation was required. The final results produced were not good enough to be used.

Did the Qualification Bodies meet all their own targets?

The short answer is "No".

They moderated the CAGs, recognising that small entry groups had to be handled differently to larger groups – moderation was applied in one of three 'levels' depending upon the historic cohort size per subject, per institution, and the 2020 number of entries for that school[2]:

No. of candidates/subject[3]	Degree of moderation for CAGs
small (reported as <5 entries)	none
medium (reported as 5-15 entries)	weighted moderation based on last three years performance
larger (>15 entrants)	overall performance largely aligned to last three years outcomes

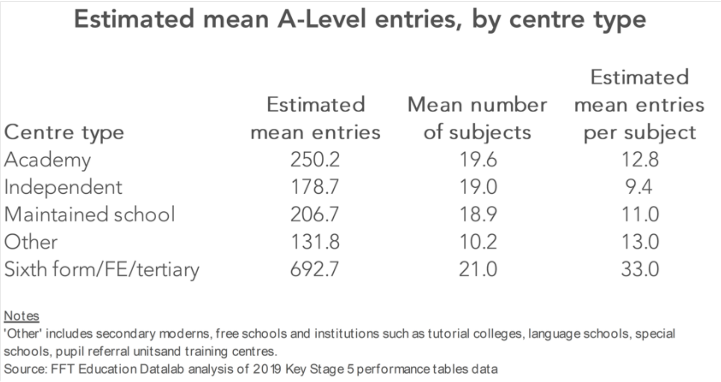
When you consider the chart below (from 2019) you can see there for smaller entry groups (which will typically be found in independent schools), CAGs would have been accepted directly or be subject to less

[1] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909124/A-level-infographic-2020-Version2.pdf

[2] The formula is quite complicated – and was reported in the press in a simplified and not entirely accurate form – I believe the details can be found in the chart found in George Stone's comment linked to the table heading below.

[3] It should be noted that cohort size calculation was calculated by the 'harmonic mean' calculated over the last three years' data. For an explanation of this you can look here: https://twitter.com/DrStoneMaths/status/1294598040628727811

moderation. 6th Form and FE Colleges were most likely to have results based on historical distributions which would be applied to the ranking rather than estimated grade in the CAG. This does not account for all the anomalies – but it represents one significant problem area.



Estimated mean A-Level entries, by centre type

Centre type	Estimated mean entries	Mean number of subjects	Estimated mean entries per subject
Academy	250.2	19.6	12.8
Independent	178.7	19.0	9.4
Maintained school	206.7	18.9	11.0
Other	131.8	10.2	13.0
Sixth form/FE/tertiary	692.7	21.0	33.0

Notes
'Other' includes secondary moderns, free schools and institutions such as tutorial colleges, language schools, special schools, pupil referral unitsand training centres.
Source: FFT Education Datalab analysis of 2019 Key Stage 5 performance tables data

[4]

Ofqual considered the impact of equality and bias on their potential models, by conducting a literature review, which informed some of their decisions. (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879605/Equality_impact_assessment_literature_review_15_April_2020.pdf).

Ofqual prioritised their macro-objectives over finding the best possible prediction for individual candidates. Professor Laura Ashe from Worcester College, Oxford described the outcome very well, when she said,

"It was clear that Ofqual felt it had to end up with a grade distribution that looked right. So they did something that made the grade distribution over the whole country look right, but they can't possibly tell us that they gave the right grades to the right people."[5]

If you are interested in a quick read on why the Ofqual moderation model was not going to achieve better than 75% accuracy at the individual candidate result level, then read Jon Coles series of tweets which can be found here: https://twitter.com/JonColes01/status/1292787888401403912

Using my summarised objectives, Ofqual initially published results provided a plausible distribution of A-Level grades compared to previous years. They produced the results on time, and arguably remained on the right side of GDPR. Unfortunately, they failed to understand the level of precision required for individual results to be credible. Consequently, they were left trying to resolve a major crisis – which resulted in their primary objective having to be sacrificed in in the interests of fairness and to avoid a

[4] https://ffteducationdatalab.org.uk/2020/08/a-level-results-2020-why-independent-schools-have-done-well-out-of-this-years-awarding-process/

[5] Guardian, August 15th, 2020

major political backlash. Well over a third of A-level entries received A/A* grades – compared to the rough average of just over 25%. The time spent on moderating the CAGs was wasted.

Why does this matter?

This demonstrates in clear sight, the consequences of poorly thought-through algorithmic decision-making operating at scale. We need to regulate the public use of algorithms, at least when they have significant impact on individuals, families or communities' lives. Where this involves public bodies, there is a chance that the democratic process might react to right perceived wrongs – but that only works when there are sufficient 'wrongs' to interest the media. Today, when technology-determined decisions have serious consequences for individuals, there is a massive asymmetry of power which leaves the subject of the process largely helpless in the face of 'the system'.

What should be done about it?

Large-scale use of algorithms must become subject to effective regulation. We need an agency (not necessarily a new one) independent of government which has statutory authority to intervene to apply appropriate sanctions in the event of misuse of algorithms. That agency needs to be appropriately resourced and staffed. Its goals should prioritise ensuring that *individuals rights and freedoms are not abused as a result of the use of algorithmic solutions*.

A minimum of three stages are required for this process to work:

1. Sensitivity Assessment. A standardised assessment of the intended use and the severity of consequences to the subject of the system when recommendations or decisions go wrong. More sensitive applications would require more scrutiny.
2. Risk Assessment appropriate to the level of sensitivity. A detailed review of the proposed application: the level of accuracy claimed [which needs to be consistent with the intended use], the evidence to justify those claims (perhaps a bit like clinical trial data), the processes surrounding the intended use of the algorithm, including the input data accuracy, decision transparency, the processes surrounding communication of the outcome to subjects, and the proposed appeal and remediation process. These need to be seen to be fit for purpose, by appropriately qualified and resourced independent assessors; to be 'safe' for public deployment. I would suggest that Algorithms should be 'licensed for use' by this regulator in the same way the MHRA/EMA licences drugs.
3. On-going monitoring and assessment. In the same way that Drug Safety processes must be in place for approved medicines; we need Algorithmic Safety processes for decision and recommender systems. These require the same level of discipline and management as those applied to Pharmaceuticals, and the regulator to be able to intervene when things go wrong.

I can already sense my fellow technologists complaining that drug approval processes are too slow for the digital age – and that is partially true. There are good reasons why those processes are meticulous.

Serious harm can be done to patients by untested new treatments. Society demands assurance that medicines are safe and effective prior to general use. The same should be true for algorithms. As they become more sophisticated and applied to more sensitive areas of life, they need better assessment and management. As with critical drugs and vaccines, there will be emergencies when decisions must be taken extremely quickly – and the regulatory environment will have to cope with that too.

We have a great opportunity to use the fiasco of predicted examination grades in 2020, to learn and change the environment to ensure the needs of individuals subject to assessment are fairly treated; not sacrificed on the altar of big-picture government targets or corporate greed.

This approach will not remove all risk from the use of algorithms to make recommendations or decisions at scale, but it will increase our 'algorithmic safety' and begin to put the needs of the algorithm's subject at the heart of any approval process which is not the case today.