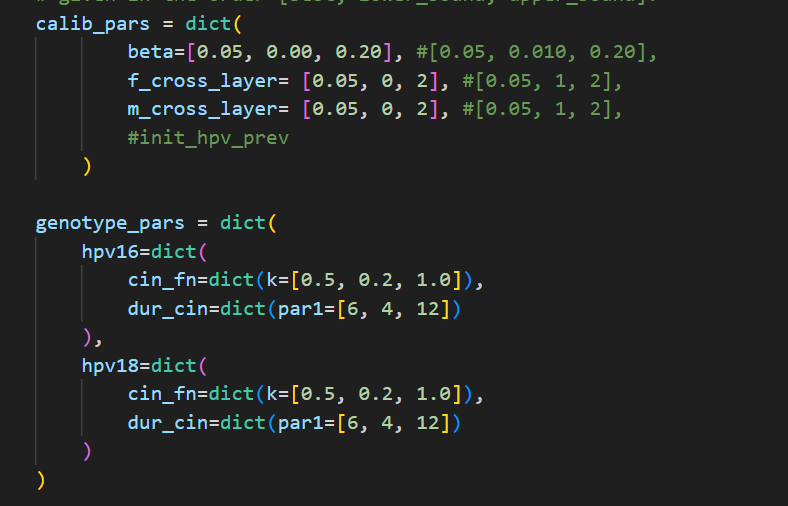
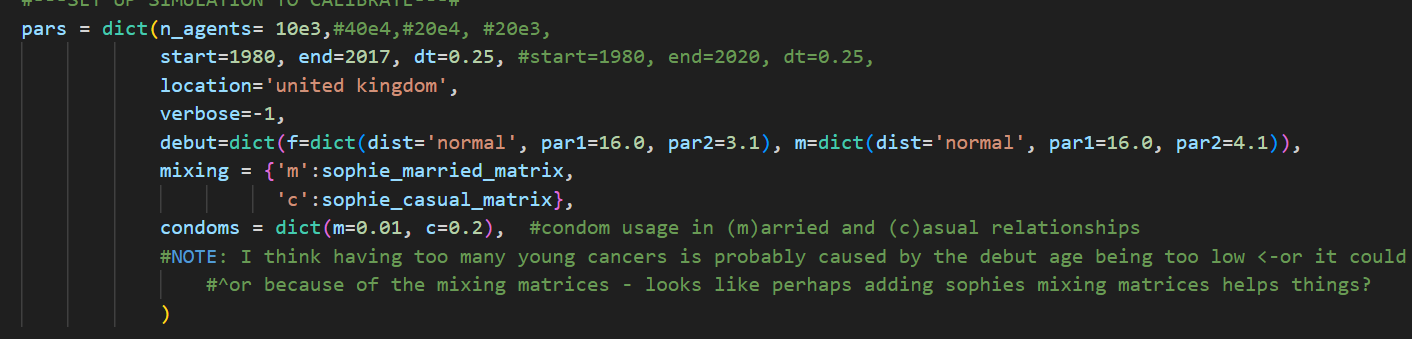
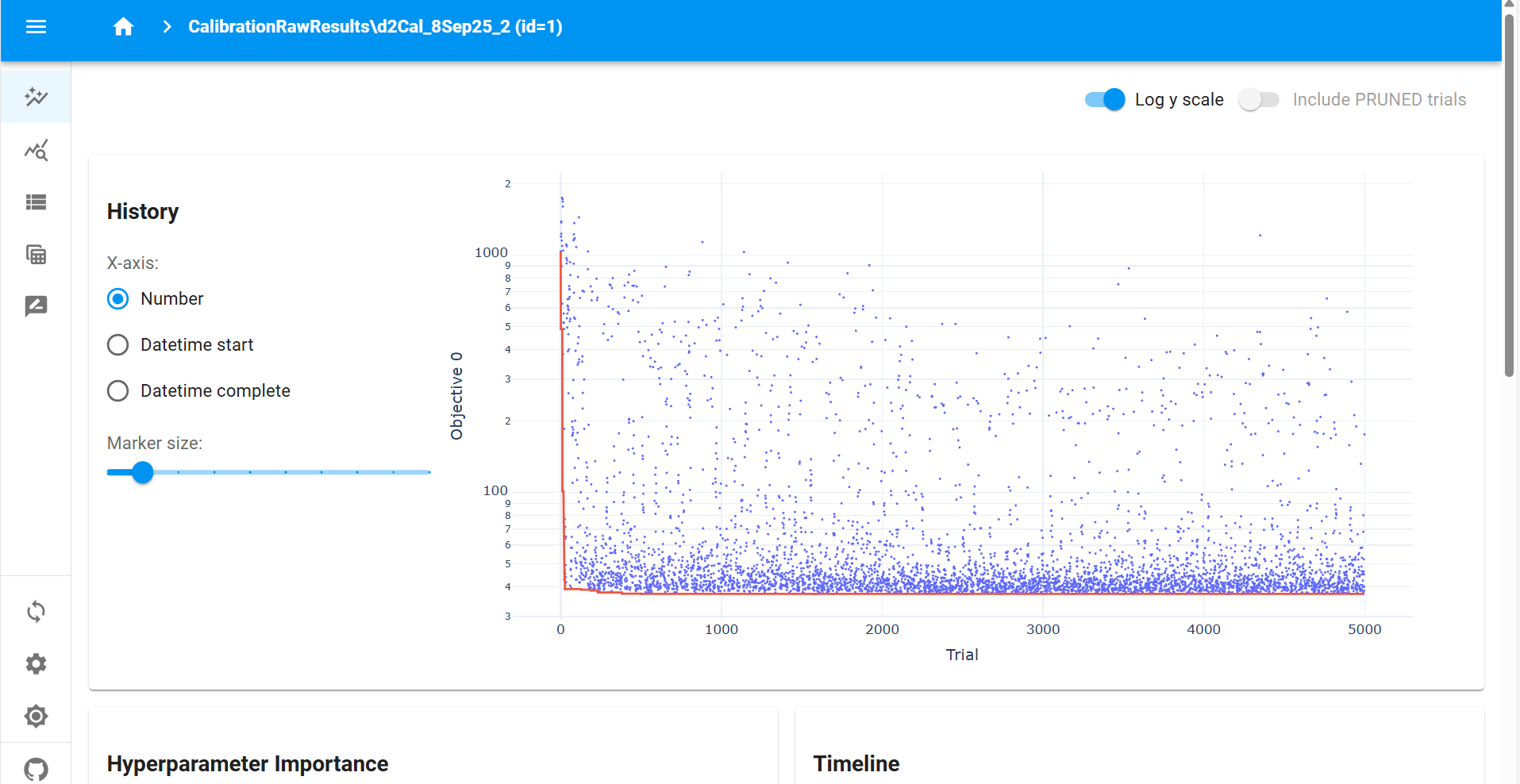
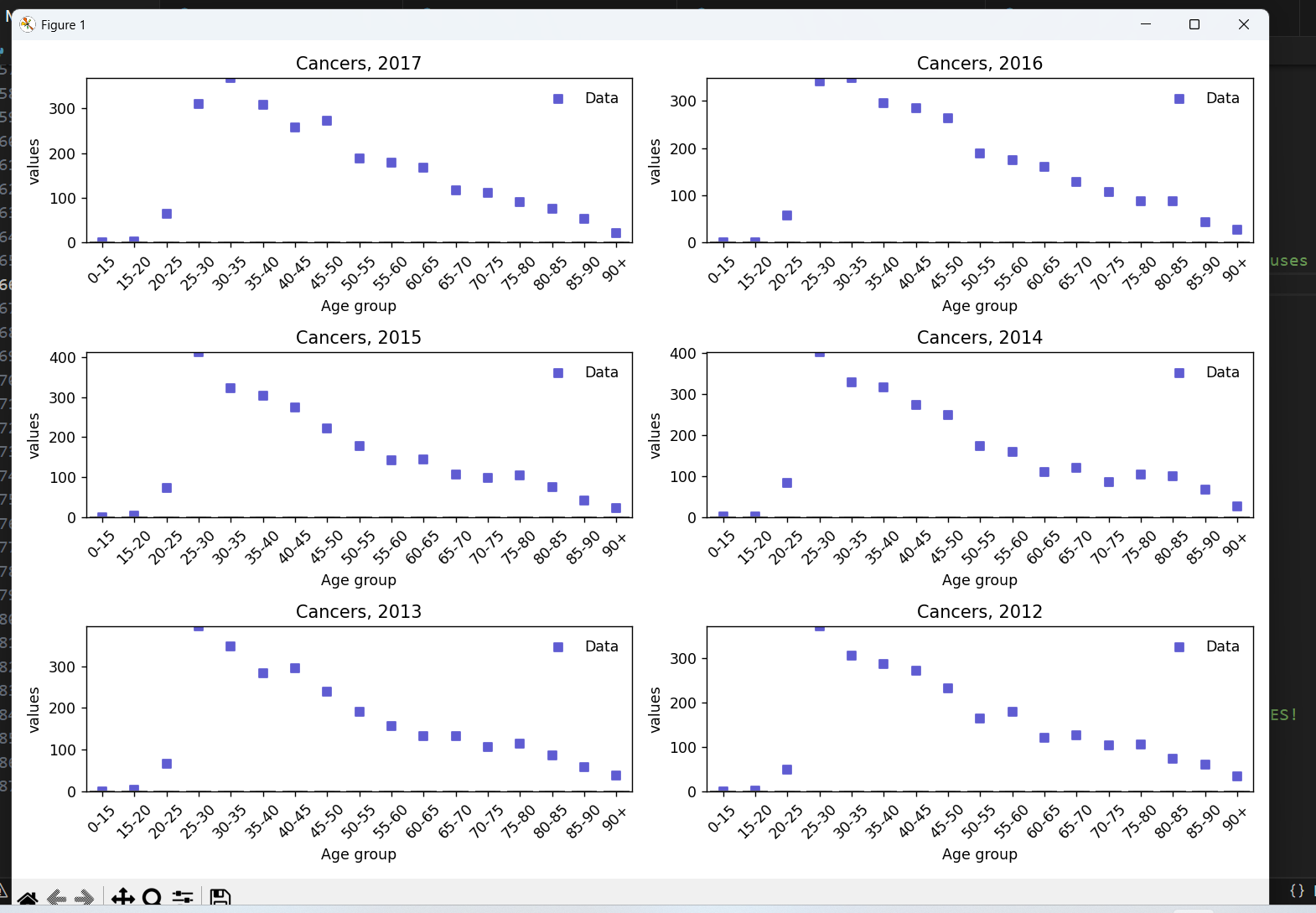
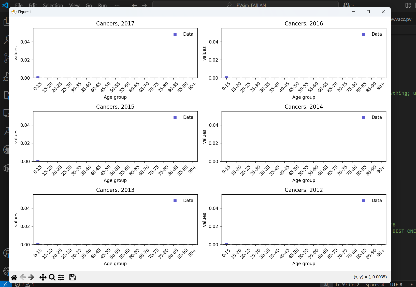
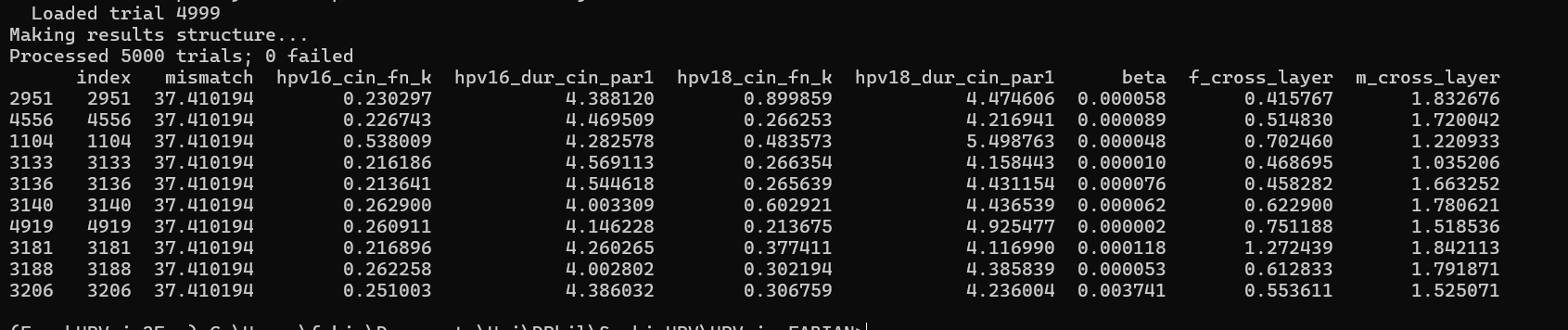
Tracking getting a UK calibration

# A

Starting off, lets try something simple and get a basic calibration to dataset 2, hopefuly decent.

**d2Cal\_8Sep25\_2**

* I want to try a calibration with the updated mixing matrices, which should allow for cancers to develop at a lower age. This should mean that I better reflect the trends in real-world data than before, where the model got a good fit for the later ages but massively underestimated how many cancers were for the lower ages
* I am not adding any of the NHS interventions right now, as it shouldn’t make a massive difference to us here (given data is 2012-2017) – of course, once I have gotten a decent calibration with no added interventions, I will model full NHS interventions to get the final cal I really want
* Only calibrating to dataset D2, so no genotype distribution: I do want to update init\_hpv\_dist to be a reasonable value (if I cant calibrate it, as calibrating it seems hard perhaps), and deffo calibrate to at least one year worth of genotype distributions
* 
* 
* It took c. 10mins to load the 5 000 trials up on Optuna Dashboard
  + 
* This calibration was utter crap. For the best 50 calibrations all the way down to the best single calibration, it is predicting absolutely 0 cancers for all age groups for all times, which of course doesn’t fit the data whatsoever
  + 
* It looks to me that the beta values for some reason are all absolutely tiny, and assuming that Optuna has done a thorough search of the parameter space (which, over 5000 trials it should have done, and it does look like it has done as we see decent variation in some other parameters – not the dur\_cin’s but i think they are all sticking rather close to 4 again to push the #cancers to 0, so that makes sense perhaps– so I think I can assume this), this means that the closest our model can get to our data is 0 cancers for each age bracket for each year
  + 
  + I think this is because there are **too few agents**. With a UK population of around 60 000 000, each of my 10 000 agents represents c.6 000 people (with cancerous agents representing 600). I suppose with the absolute values I am looking to fit to being around 10-350ish with a mean of around 150, it means the models with the best fit are just those which always predict 0 cancers.
    - If true, this further means that for dataset D2, a mismatch of 37.410194 means all model predictions are at 0 – so any lower mismatch should mean the model is doing something somewhat meaningful when fitting to the data!

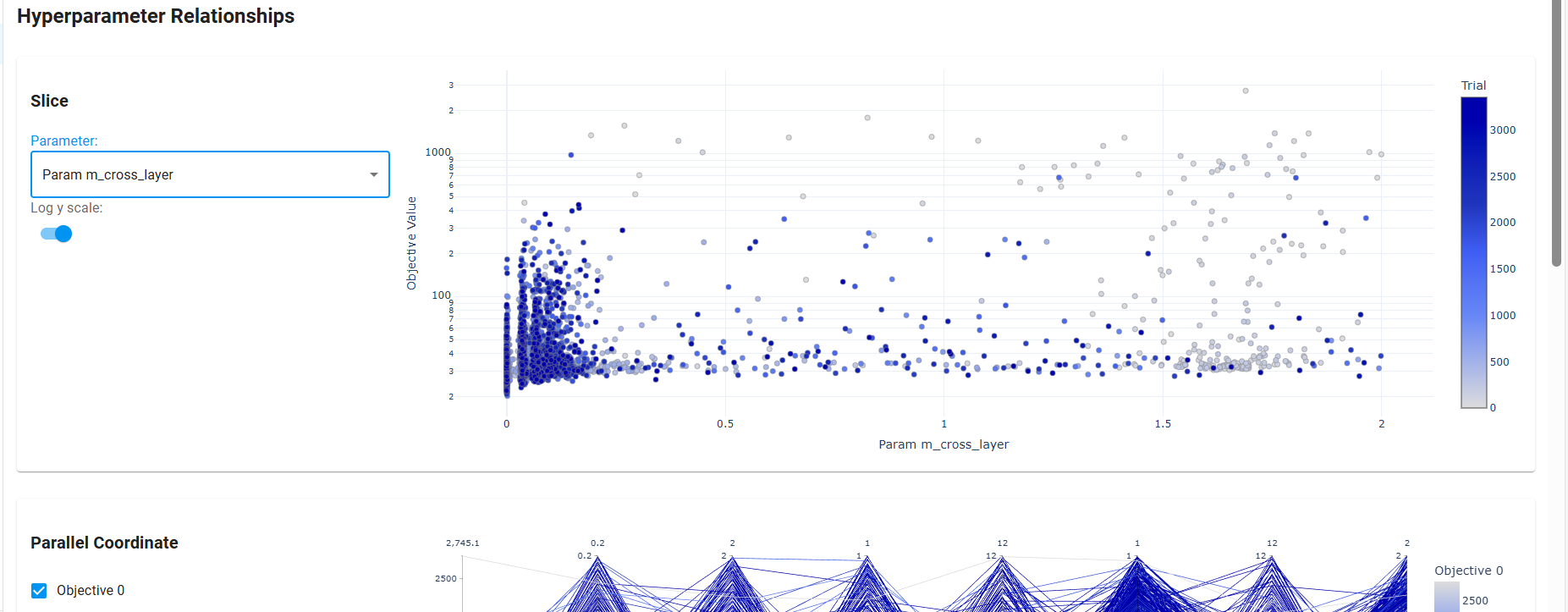
# B

Trying more agents to see if that improves things

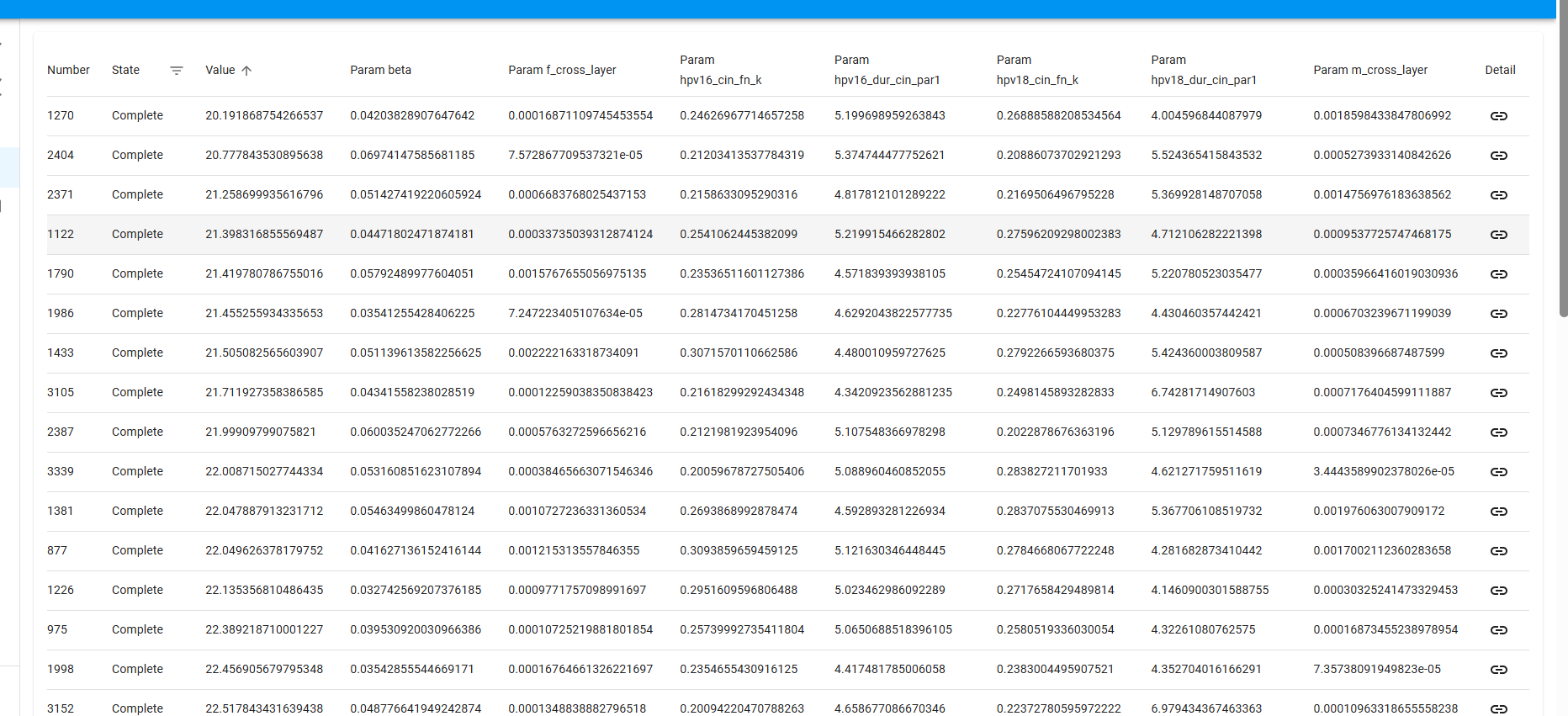
**d2Cal\_8Sep25\_5, d2Cal\_8Sep25\_6**

* I am redoing the same as **d2Cal\_8Sep25\_2** with the only difference being I am using 100e3 agents rather than 10e3 agents (i.e. 100 000 rather than 10 000). If my reasoning for why **d2Cal\_8Sep25\_2** didn’twork well is correct, then this should mean I have enough granularity with cancerous agents representing around 60 people to at least get somewhat of a good fit to the data, very crudely.
  + If this is true, it is then time to either try a calibration with even more agents (maybe x4 so each cancerous agent represents around 15 people), or by grouping some of the age brackets for the data (perhaps into 10-year buckets), or both, to try and get a model which fits the data really nicely.

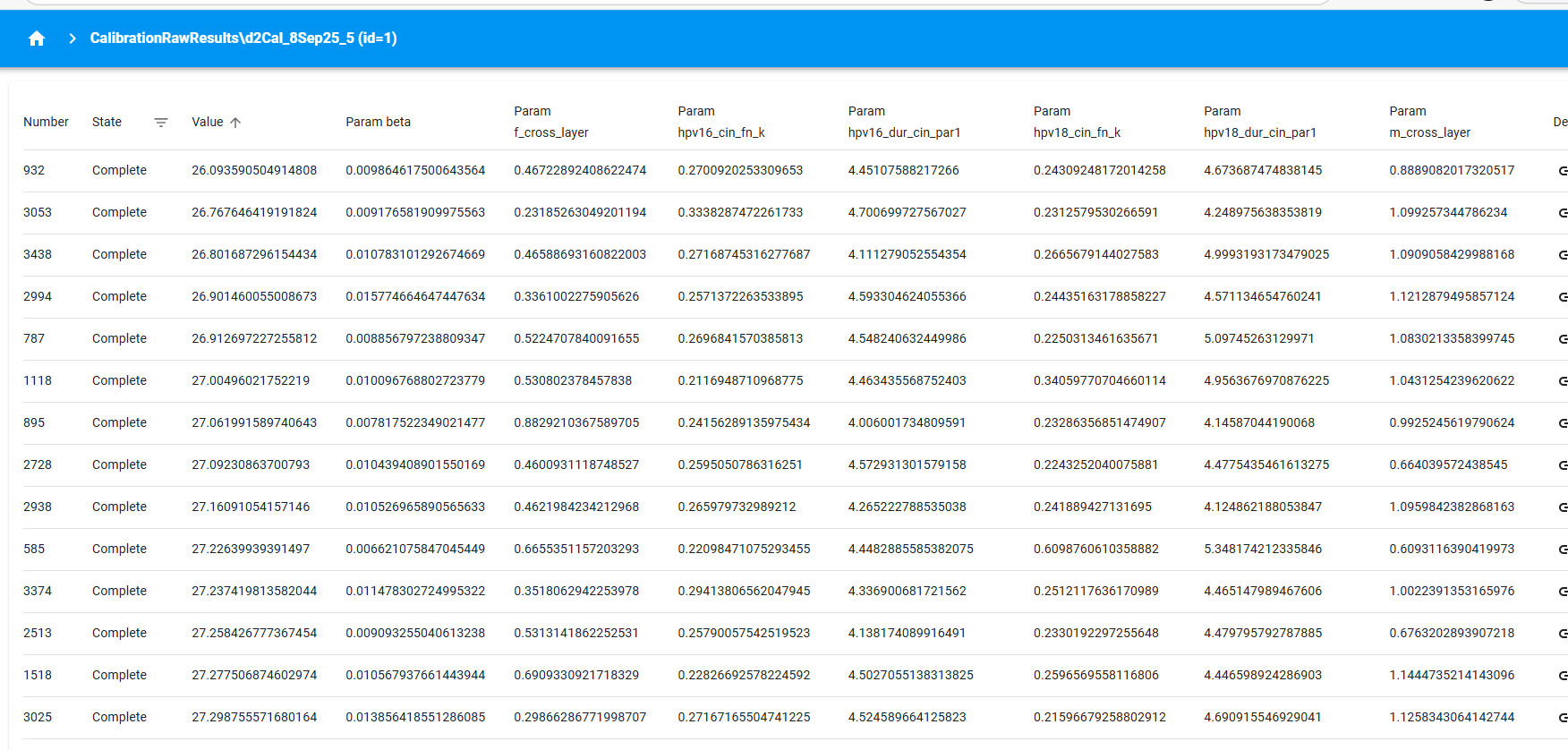
….. i think that maybe as trials continue, more data is stored up and maybe my computer ran out of memory and that is why it crashed at 4000 trials.. if so it is no matter as i can still load up the study in Optuna dashboard and get the best parameter values on there and rerun with these values and see the tightness of the fit. And either way, i should still use that calibration therefore! Even if the rerun somehow works because i have done die=False rather than die=True



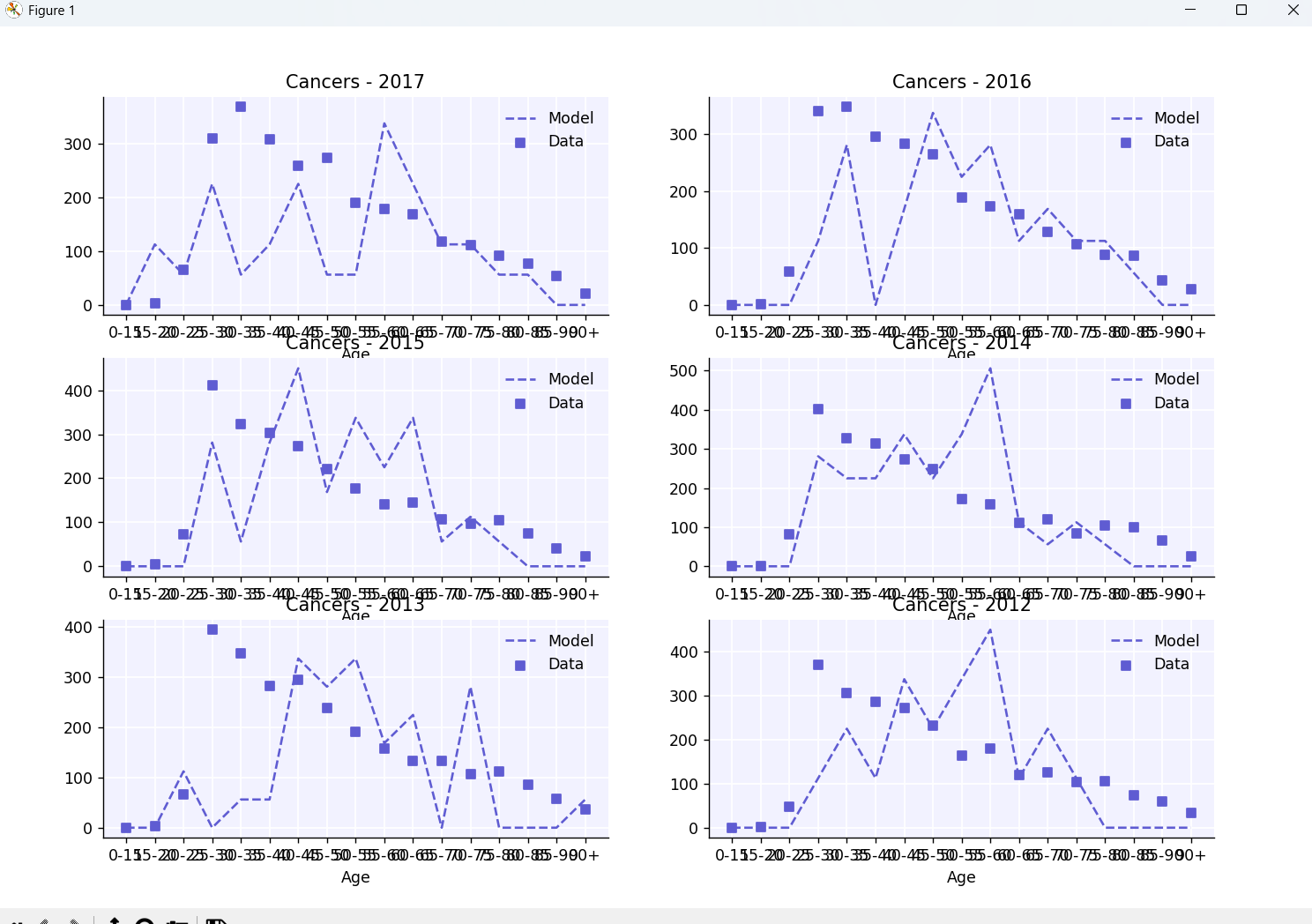
^its worth looking at this i think to see if i need to make ranges of stuff bigger , i think in cases like this where stuff is bunched to the side, probably yes!.



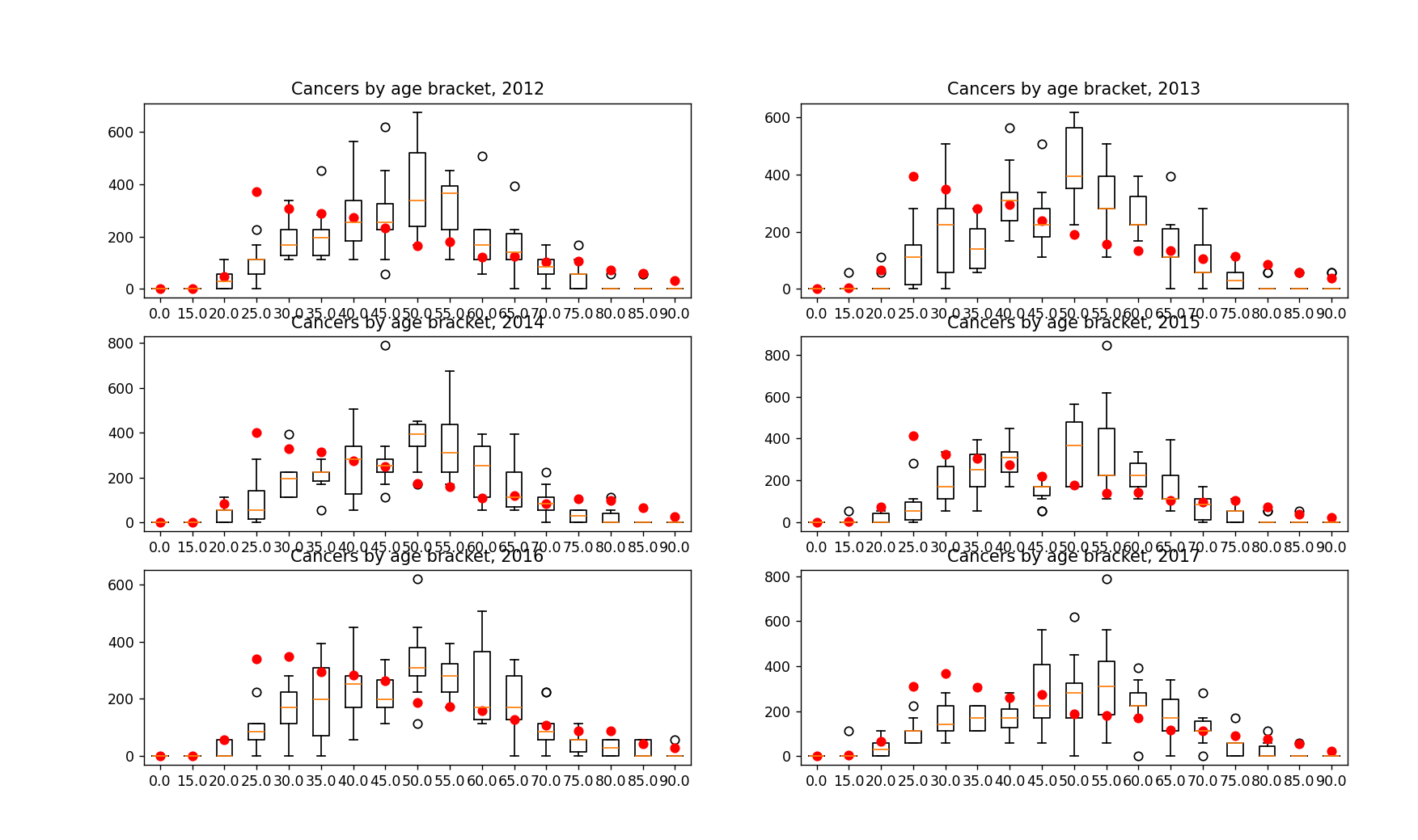
^ for number 6



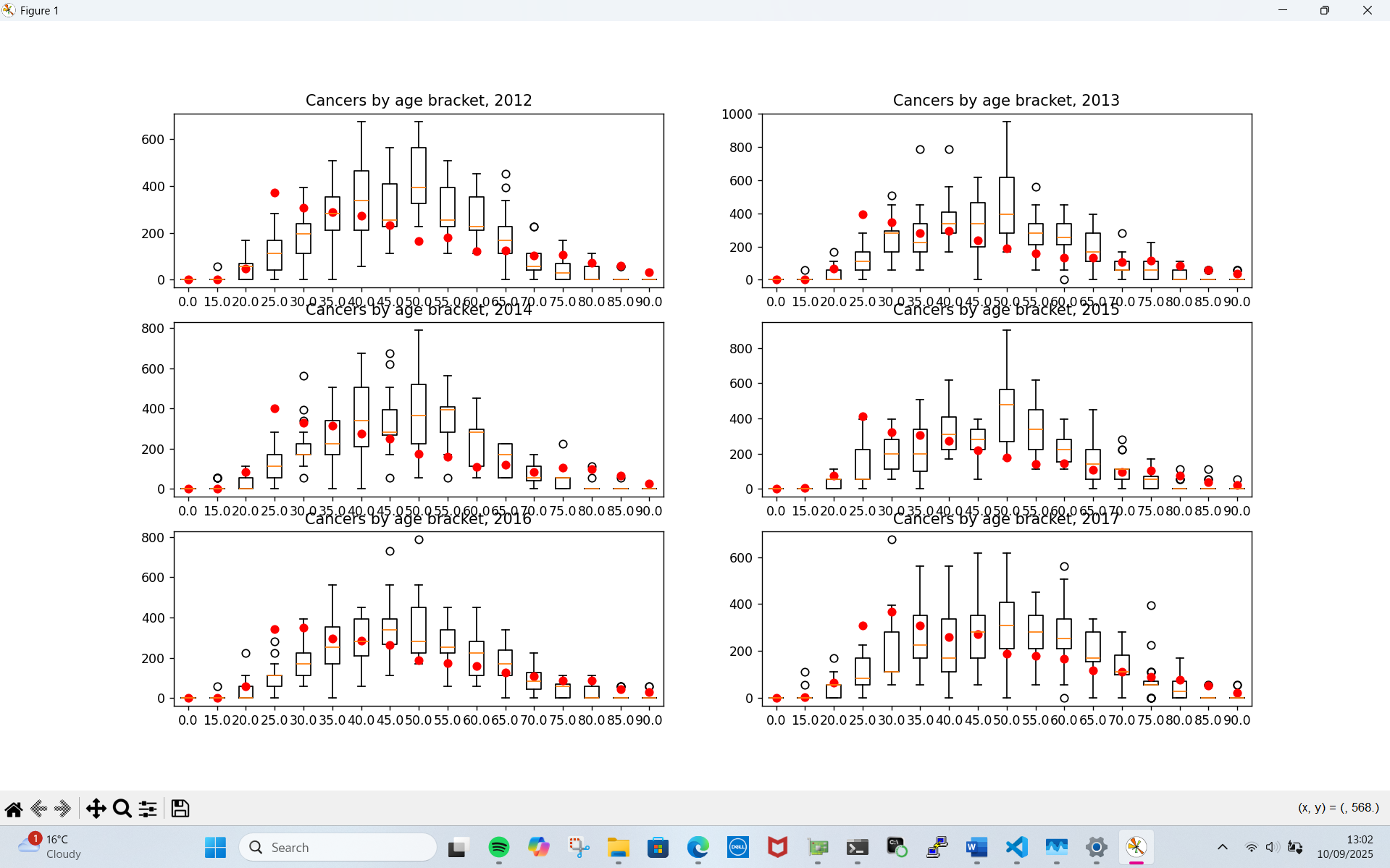
^for number 5



^ one run from best of number 6



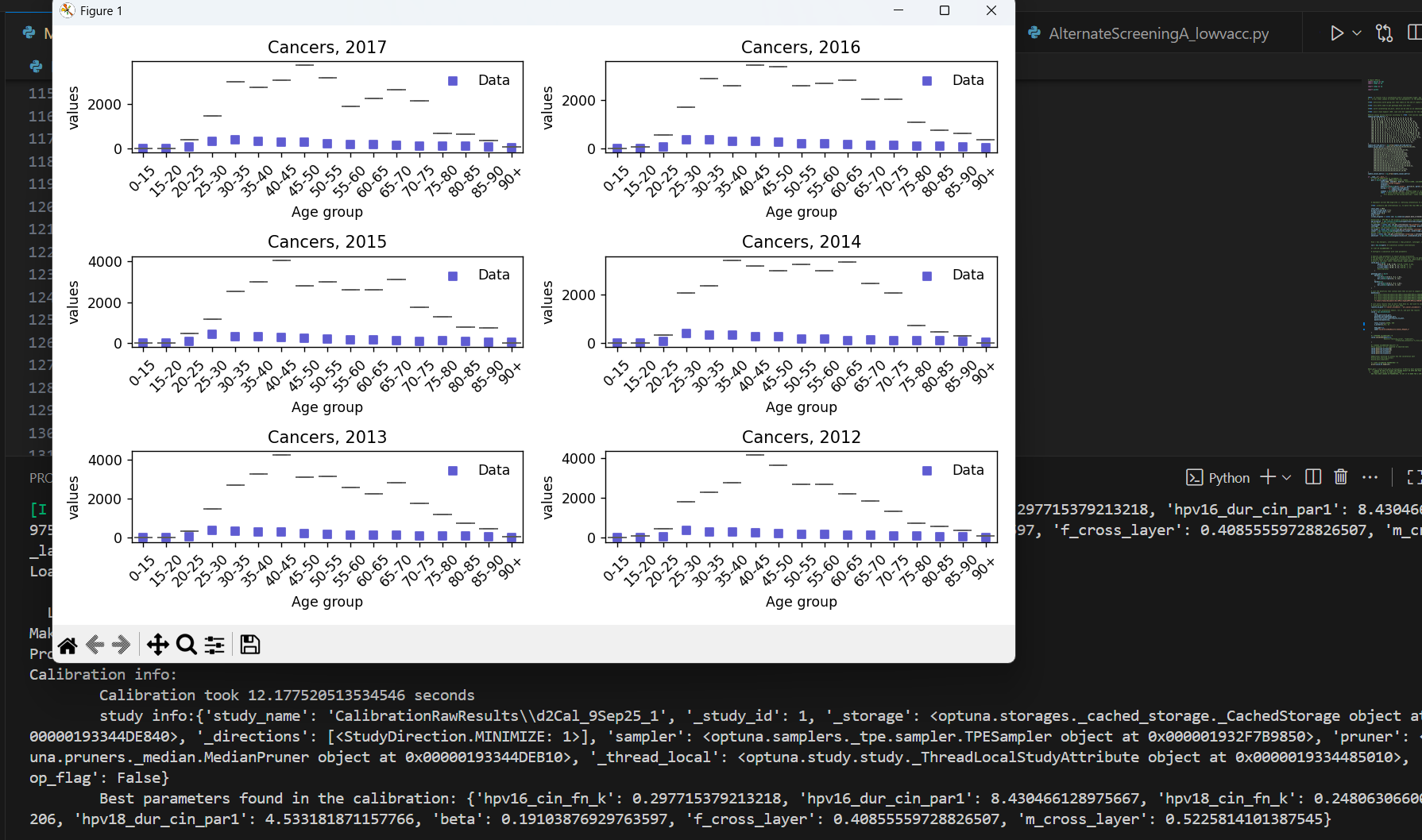
^ 10 runs from the single best of number 6



^ 5 runs for each of the 4 best fits of cal 6

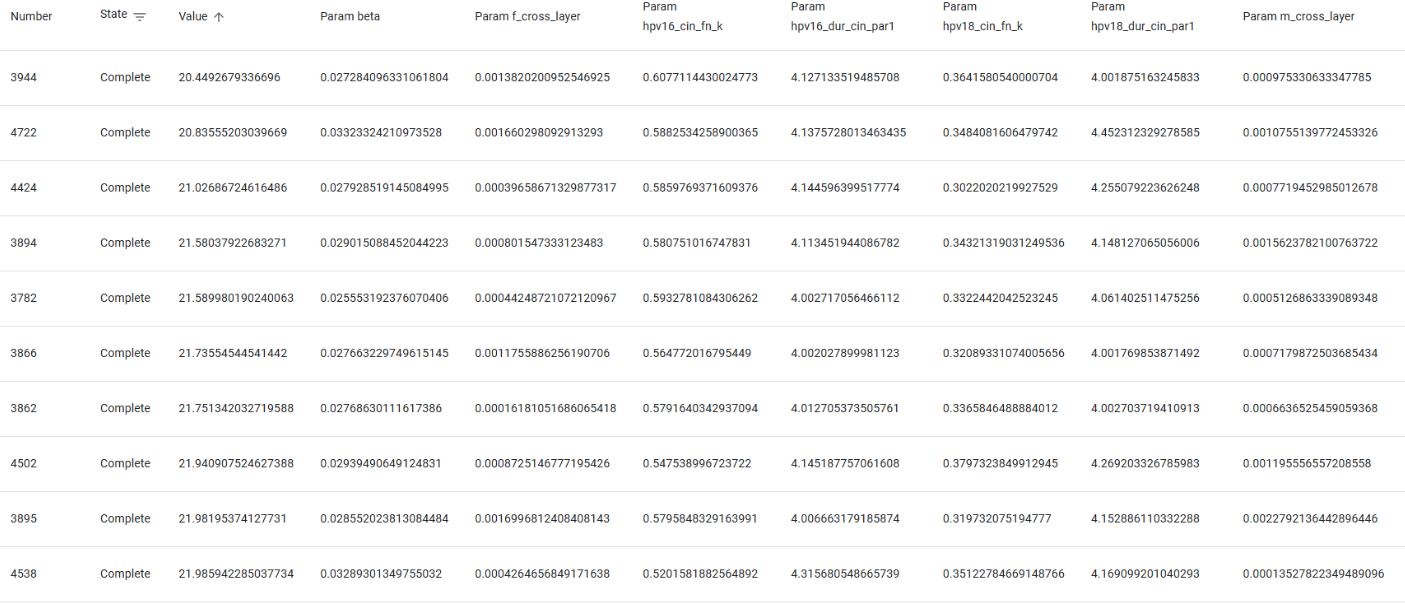
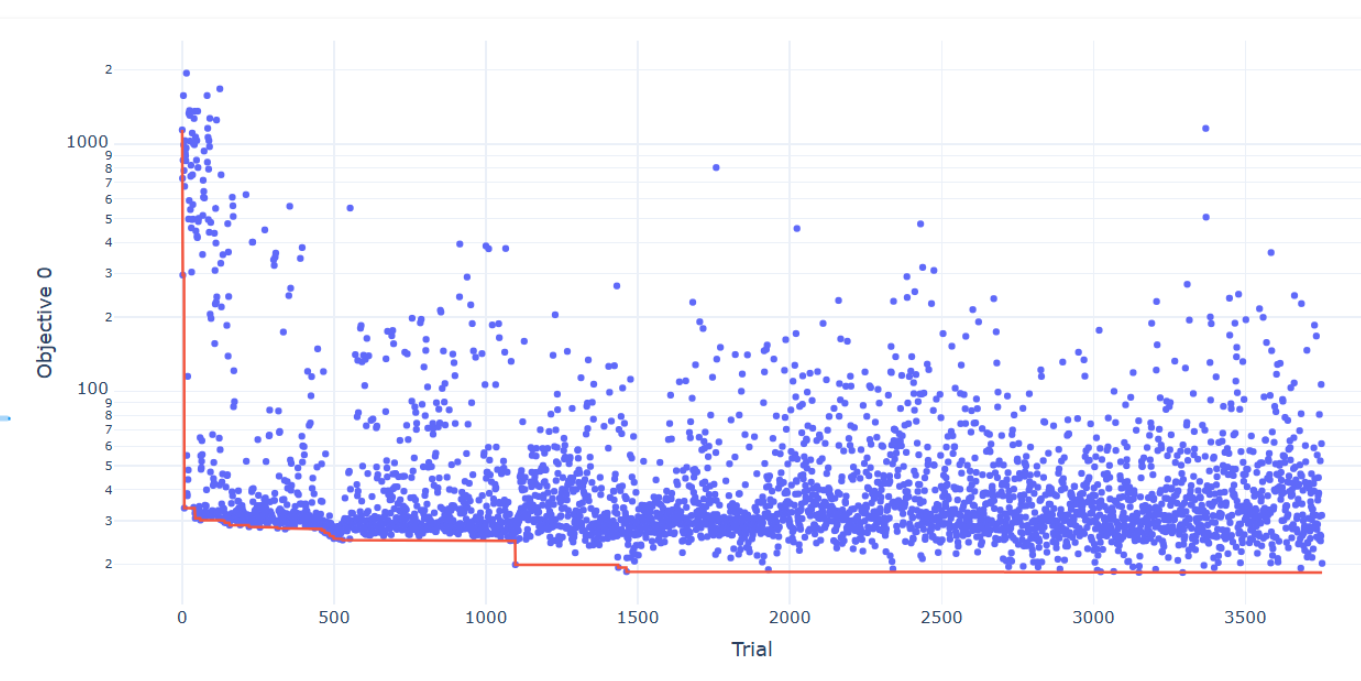
**d2Cal\_9Sep25\_1**

interestingly, when running a single trial cal i found:

* Optuna does not start off with our ‘best guesses’ for parameters, meaning I can’t use that to control the parameters used in a calibration, to retry them
* With the current config (as in **d2Cal\_8Sep25\_5, d2Cal\_8Sep25\_6** ) I am able to get cancers at a sufficiently low age, at least when beta is big enough (noted **d2Cal\_9Sep25\_2** has much smaller beta and no cancers below a certain age, while **d2Cal\_9Sep25\_3** hasa bigger beta of 0.01 at least and does have cancers below a certain age, alebit fewer. I think its fair to blame the differences on the beta)
* 

# C

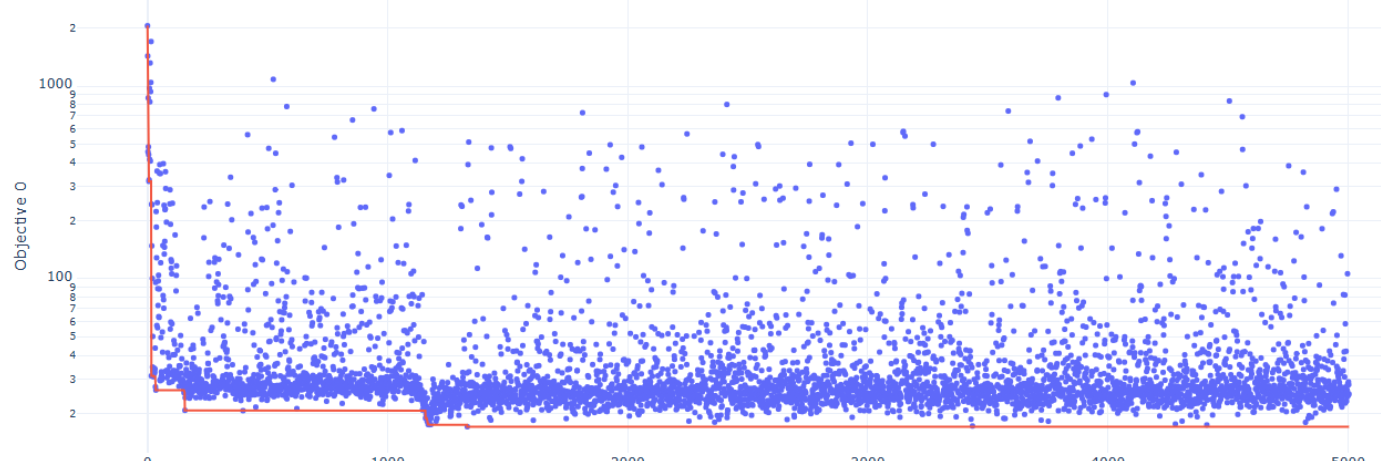
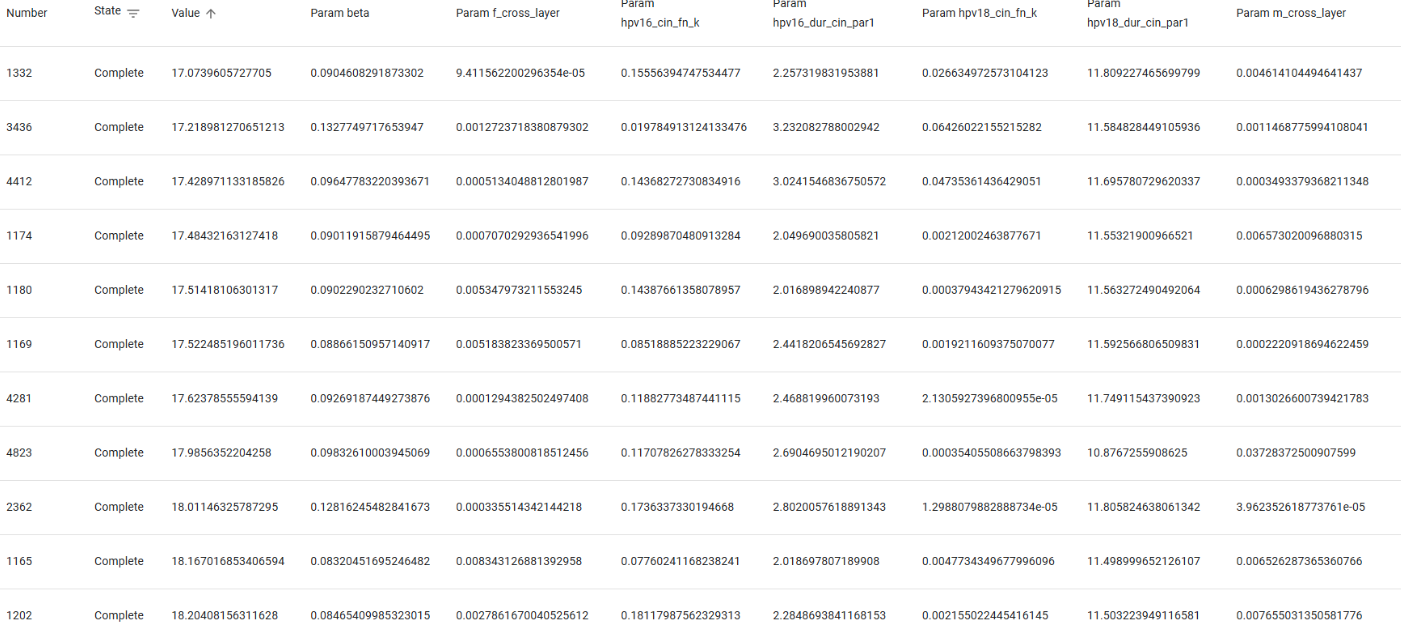
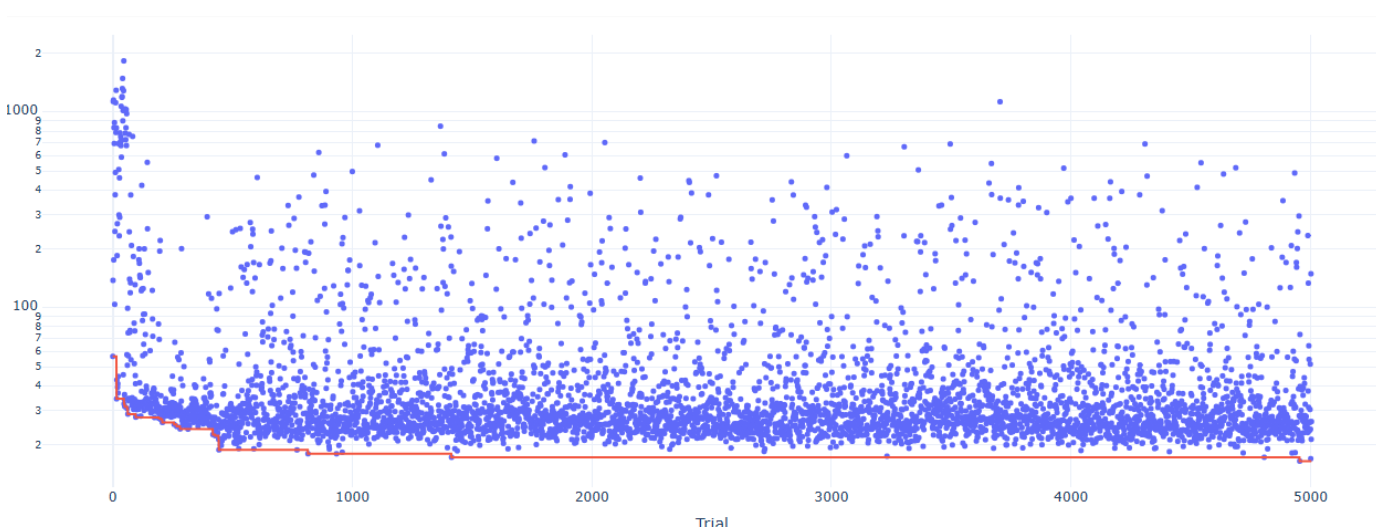
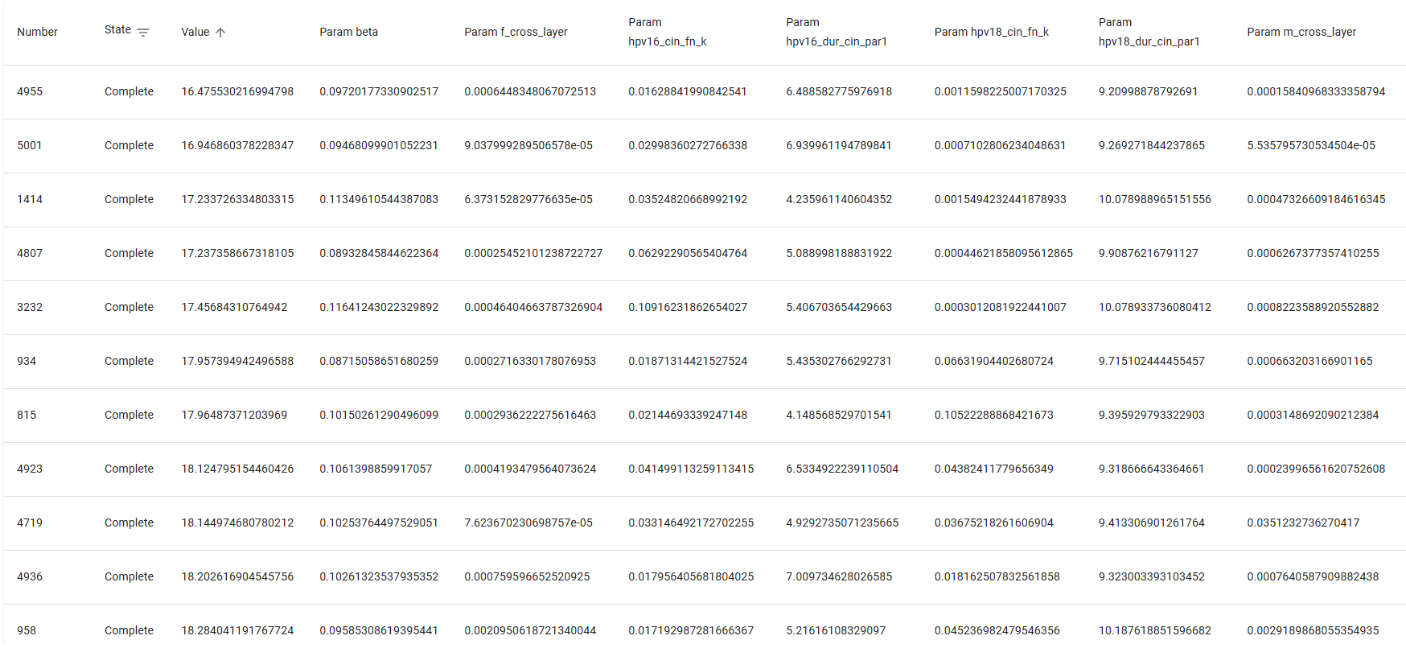
**d2Cal\_10Sep25\_1, d2Cal\_10Sep25\_2**

* It does look better with 100 000 agents, so now I am trying out 250 000 agents.
* As with 100 000 agents we seem to capturing the right level of granularity, i am not expecting a serious improvement if any
* I am running the cal twice to get some robustness against the randomness of the calibration process itself
* In both cals, we don’t see a noteworthy improvement upon our B calibrations
  + Each full trial of the C cals takes 2.5x as long as a trial of the B cals, and they require 2.5x as much memory at their max memory need, which means I can do many fewer in parelell, meaning we get a double whammy in slowdown.
  + … but if it is not signficiantly better than B, it is not worth the slowdown because I can do several B cals in the time of a C cal, and with the noise in calibration results, often trying many identical-set-up cals gets better results than one on-average-better cal that takes the same time
  + [also, given that extending the ranges as in D also gets very good results, we really cant justify increasing to 250 000 agents from 100 000]
* **d2Cal\_10Sep25\_1**
  + 
  + 
* **d2Cal\_10Sep25\_2**
  + 
  + 

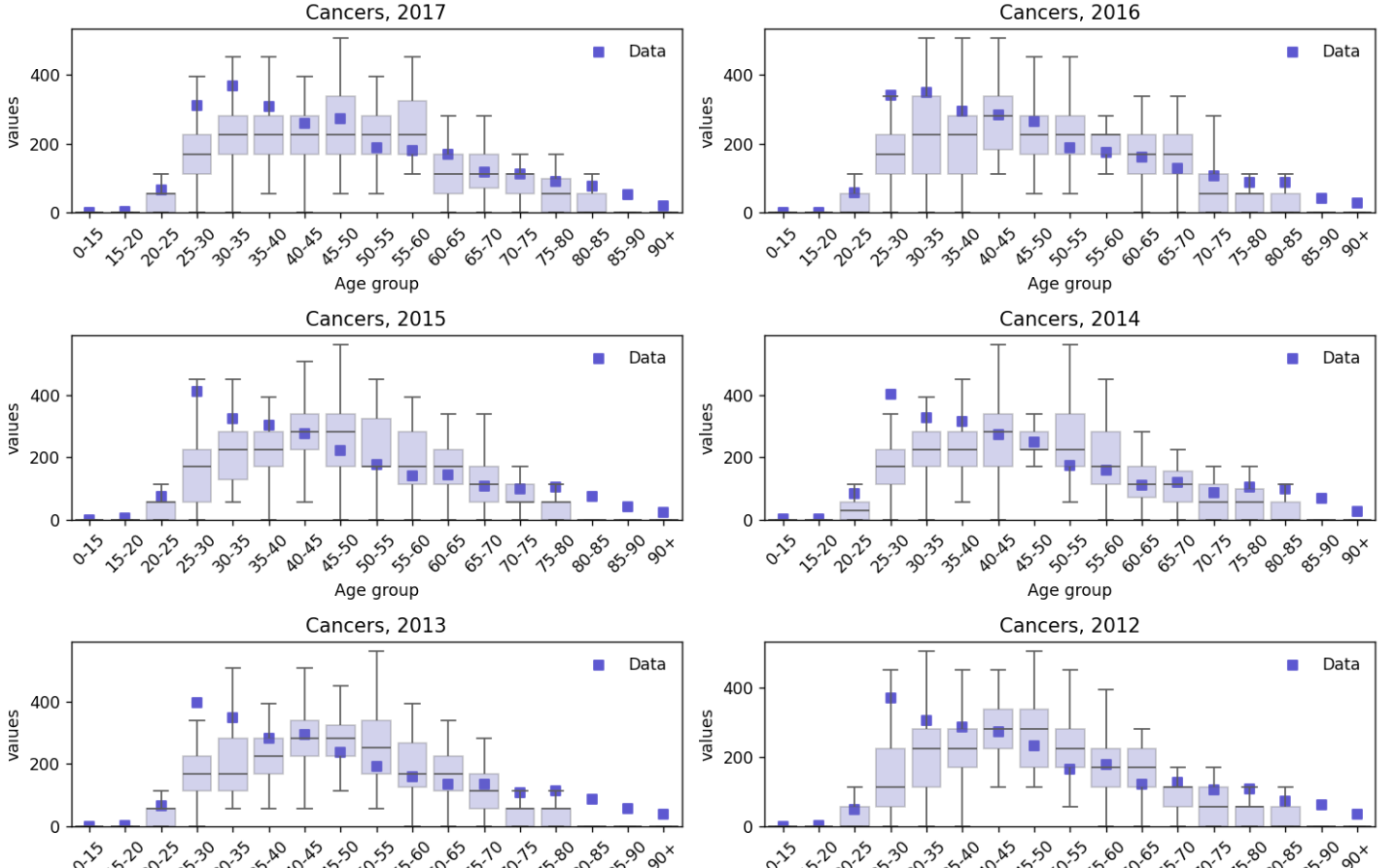
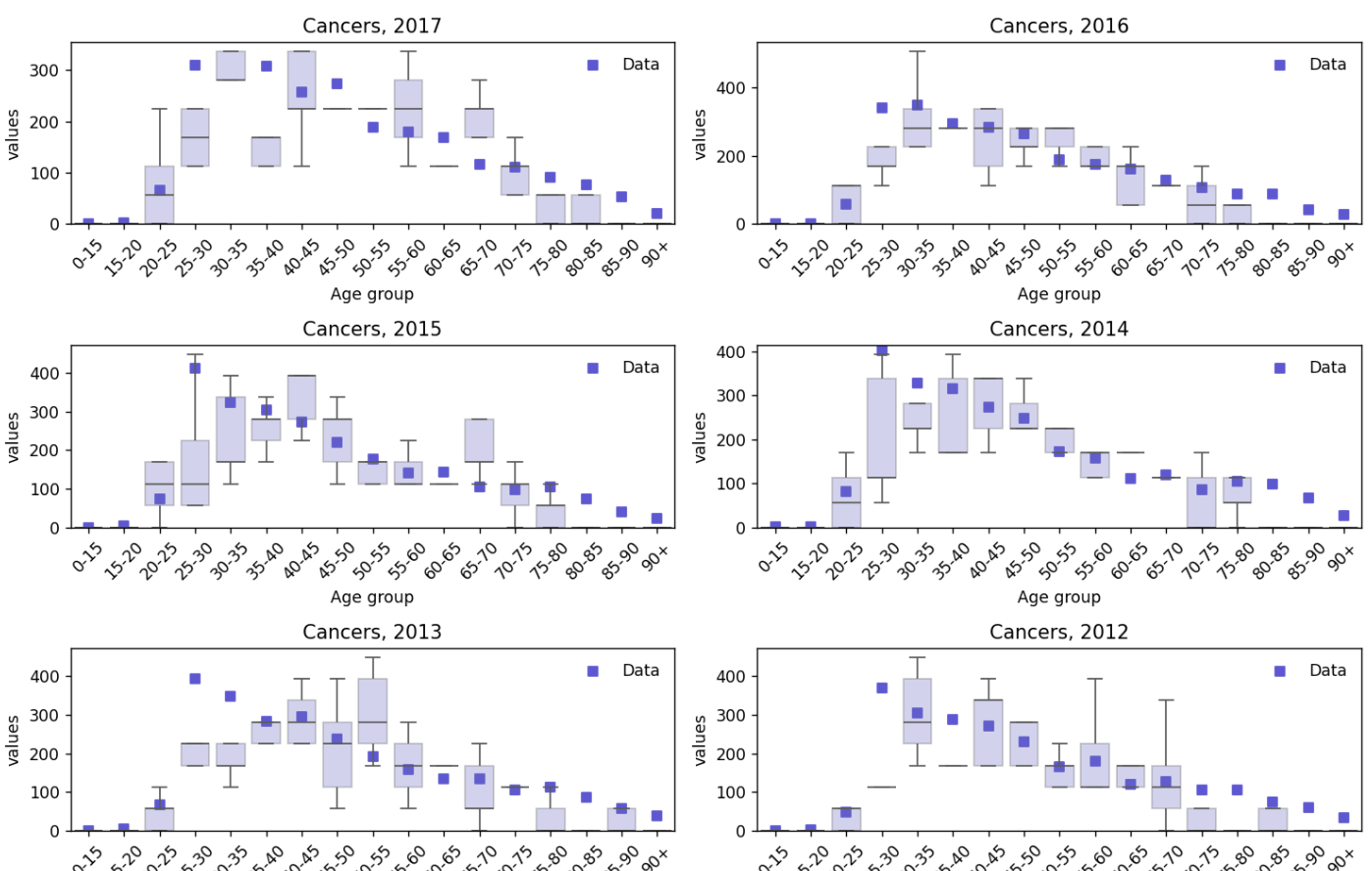
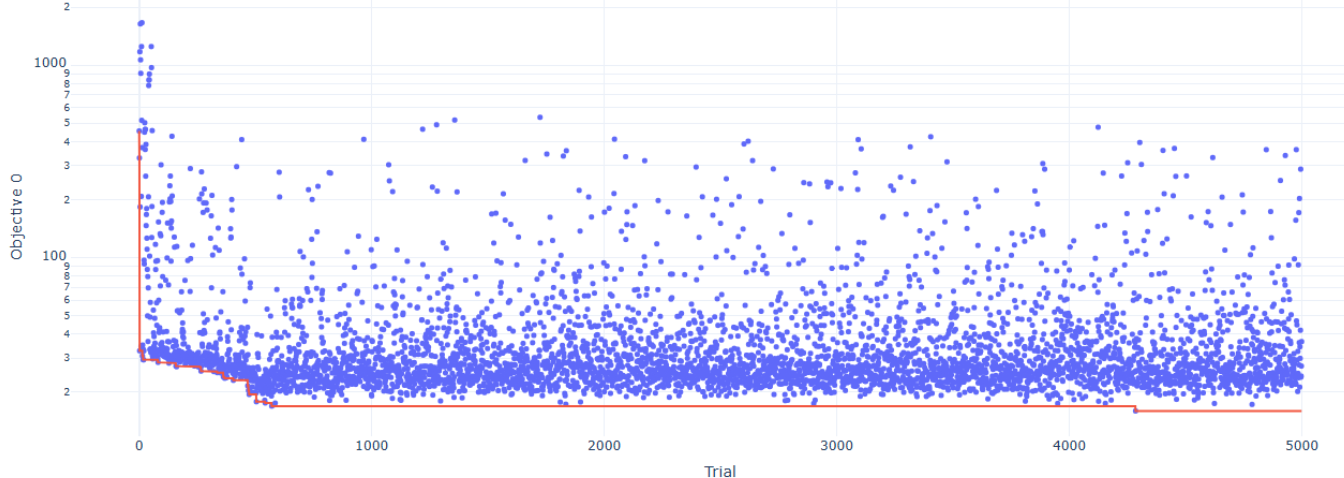
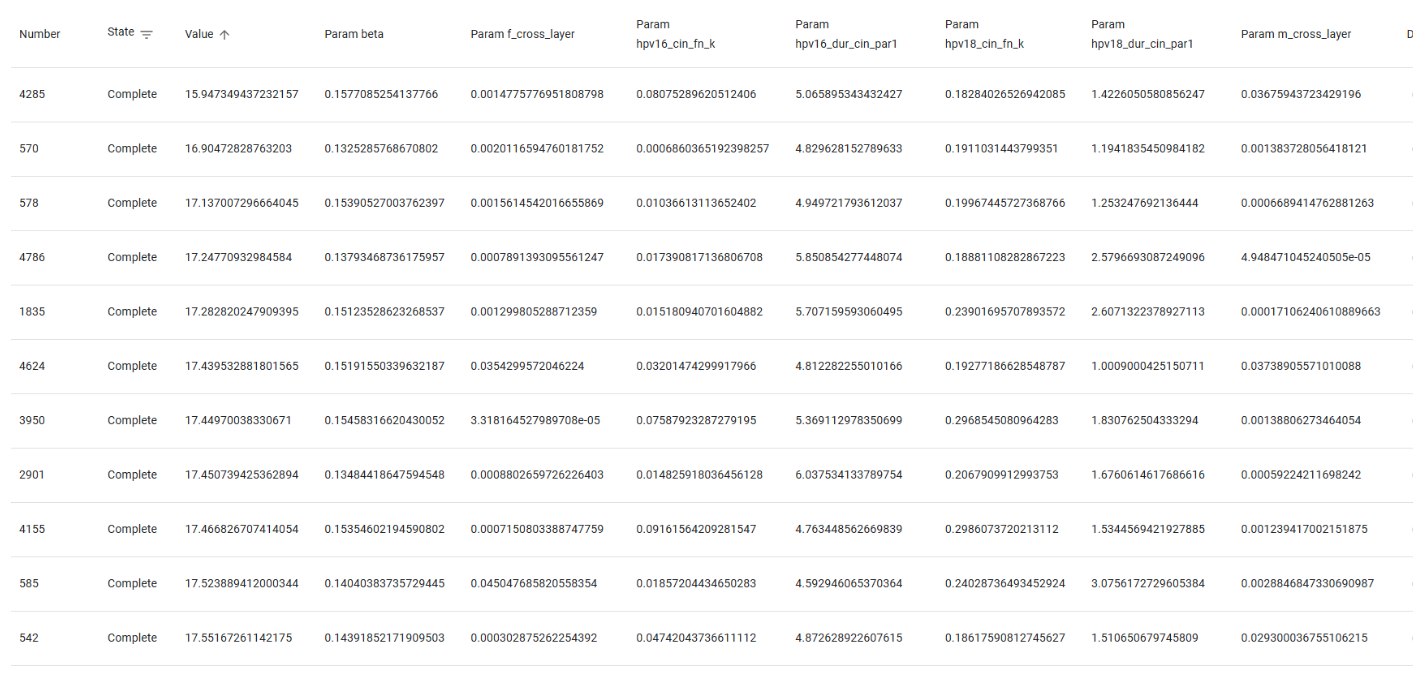
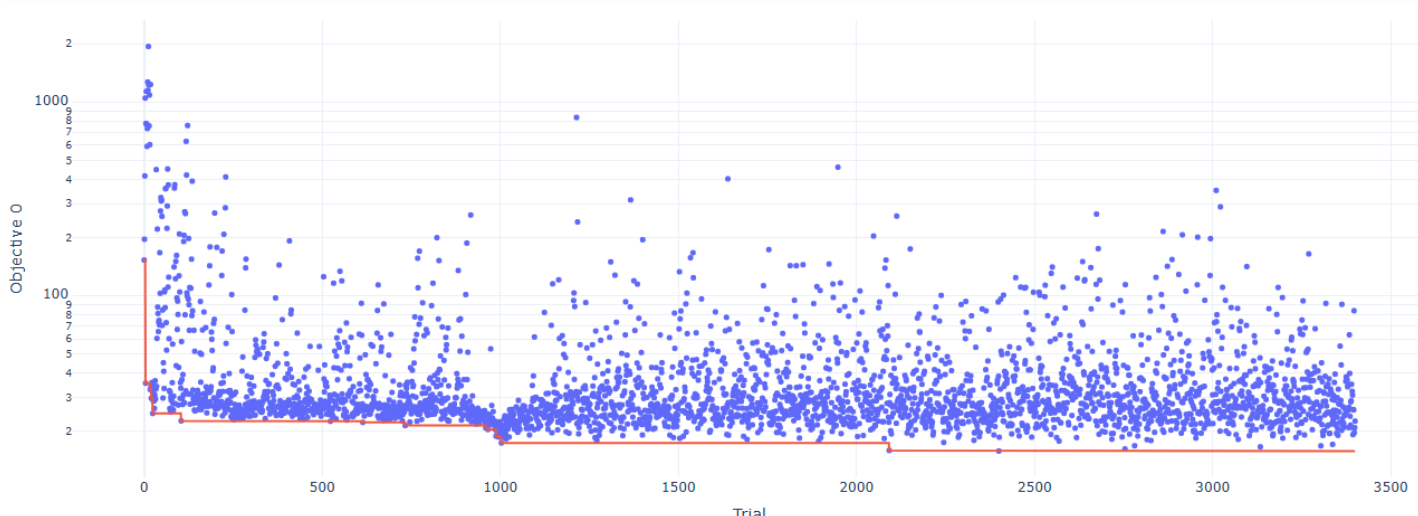
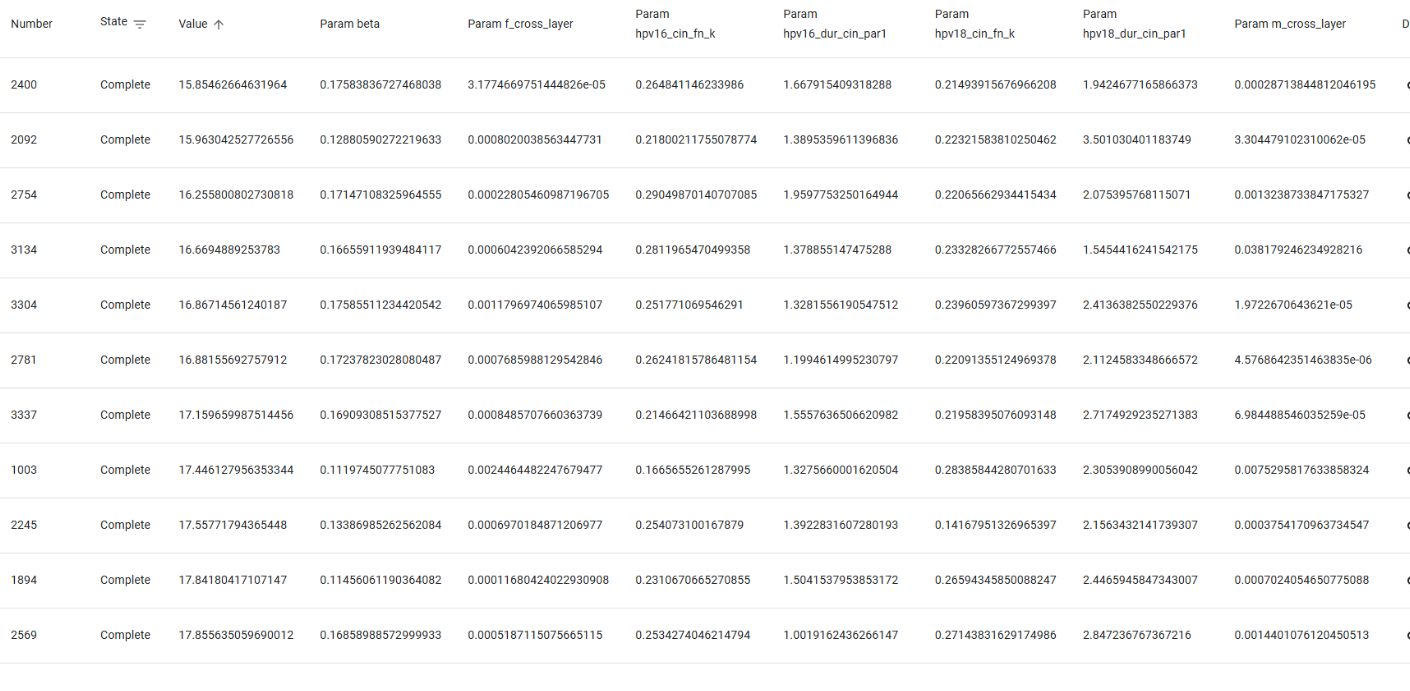
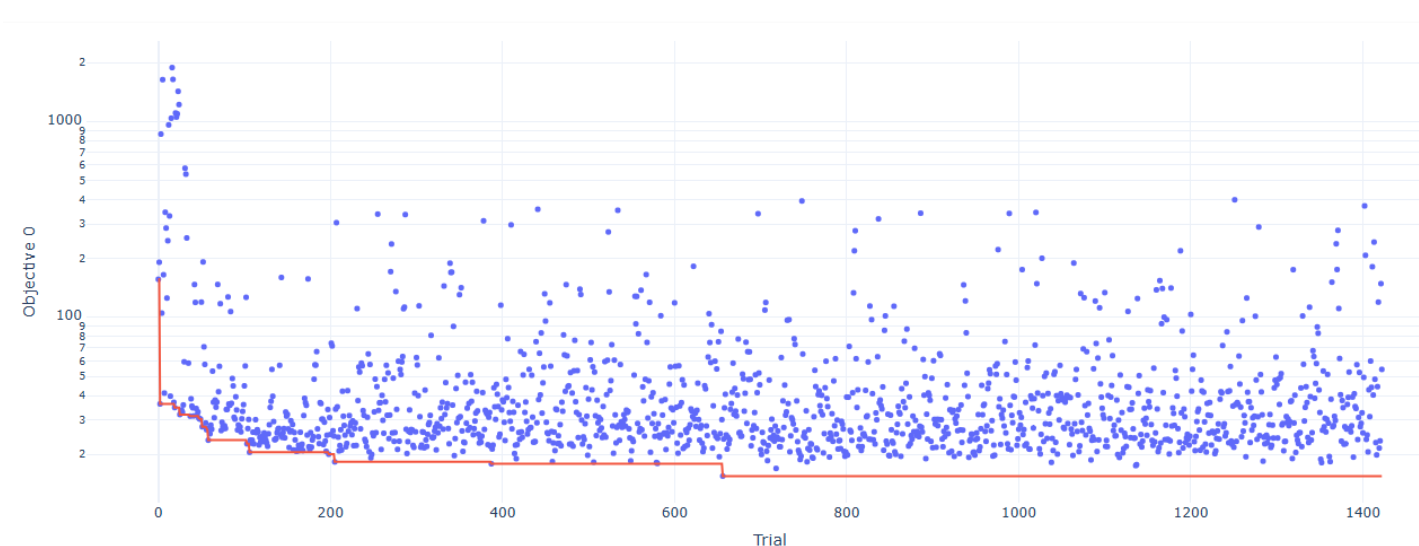
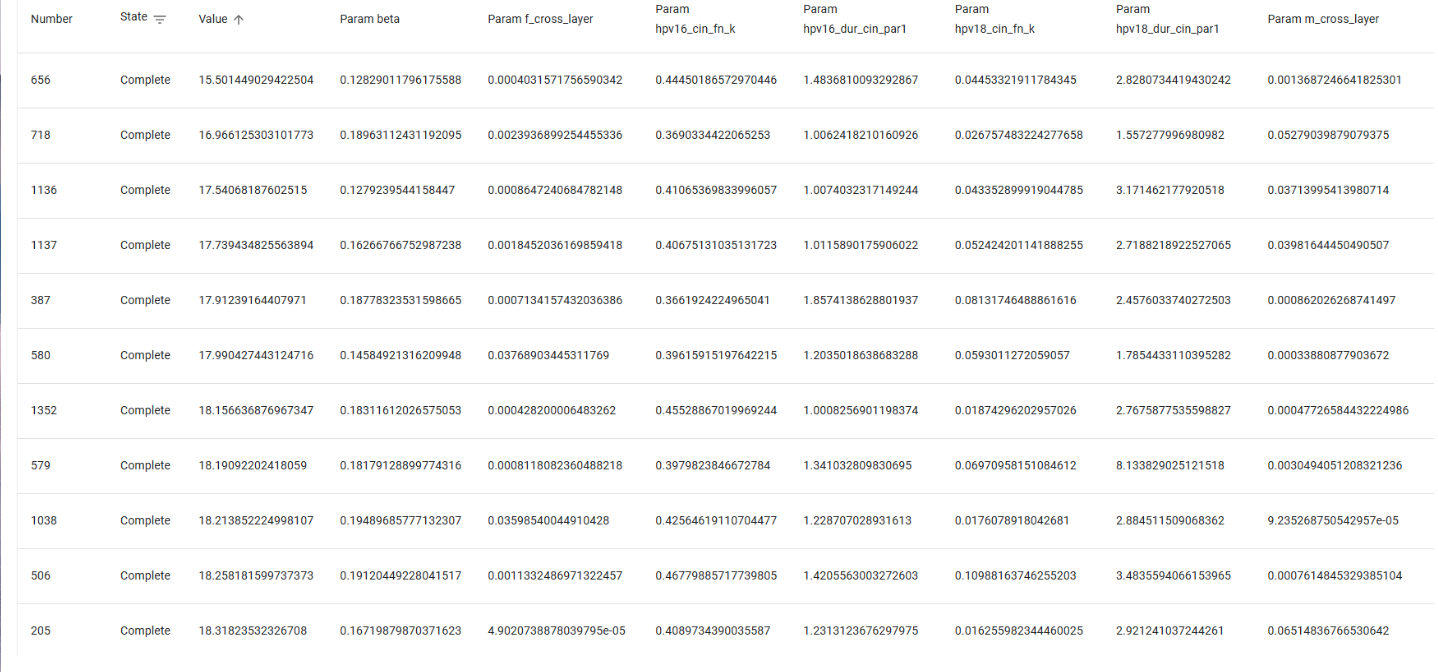
# D

A tricky thing here is that different calibrations may end up exploiting different parts of the parameter space and settle on discovering different parts of the parameter space as the best regions, maybe getting as good GOFs as each other but with some parameters really differing. This makes a difference and means I shouldn’t be too hasty in reducing parameter ranges, but does mean I should think more about increasing them at least. Maybe reducing sometimes tho.

**d2Cal\_11Sep25\_ZB\_D2, d2Cal\_11Sep25\_ZB\_D3**

* Calibrating (ZB means on zenbook) identical to **d2Cal\_8Sep25\_5, d2Cal\_8Sep25\_6**, but with extended parameter ranges (CHANGES: 16\_cin\_fn [0.5,0,1], 16\_dur\_cin [6,2,12] and 18\_cin\_fn [0.5,0,1], 18\_dur\_cin [6,3,12])
* **d2Cal\_11Sep25\_ZB\_D2**
  + ****
  + ****
* **d2Cal\_11Sep25\_ZB\_D3**
  + ****
  + ****
* Great news – 100 000 agents with an extended range does seem to consistently get better results than even 250 000 with the old ranges. Nice!
* Looking at these two cals to see if I can extend the range even further, it looks like I could.

**d2Cal\_12Sep25\_ZB\_D4, d2Cal\_12Sep25\_XPS\_D5, d2Cal\_12Sep25\_XPS\_D6**

* CHANGES TO RANGES COMPARED TO A,B,C: 16\_cin\_fn [0.5,0,1], 16\_dur\_cin [6,1,12] and 18\_cin\_fn [0.5,0,1], 18\_dur\_cin [6,1,12]
* Hoping to see at least as good cals as before (that is, after 5000 trials under 20 consistently), if we can get even better, that would be great. (Then I am ready to improve the fixed parameters of the HPVsim model and carefully pick which parameters will get calibrated in the final cals, that is, doing E)
* **d2Cal\_12Sep25\_ZB\_D4**
  + best 50
    - 
  + Best 5
    - 
  + 
  + 
* **d2Cal\_12Sep25\_XPS\_D5**
  + ****
  + ****
* **d2Cal\_12Sep25\_XPS\_D6**
  + ****
  + ****
* Well, that worked well! By extending the ranges, I am consistently getting below 20 and even the lowest GOFs of all of them, and some cals which do look seriously not too bad! Time to carry this over to step E

**I have stopped editing the code as used in these calibrations in the D section, so that if things all go wrong in the future with E, I can go back to something that deffo works well – and this code is called *MYcalibrationD.py*.**

# E

Note: before, it looks like sophie just did the genotype parameters calibrated exactly as in the tutorial, so here I am carefully going over all HPVsim parameters and either (A) confirming I am happy with the values they are right now, (B) finding literature-guided new values for them for the UK, (C) determining that they need to be calibrated.

**Parameters of HPVsim that I am looking at – these are figuring out parameters within hpvsim (not, for example, efficacies of treatments and stuff which are other parameters I do need to look at, but elsewhere – in the NHS\_Screening\_Pathway.py file instead):**

* *init\_hpv\_prev*,  *rel\_init\_prev, init\_hpv\_dist* <- for a sensitivity analysis
  + init\_hpv\_prev gives the initial probabilities of someone having hpv, given they are sexually active, when setting up the simulation.
    - I am informing this better with <https://hpvcentre.net/statistics/reports/GBR.pdf?t=1757520103139> (that is, that Human Pappilomavirus and Related Diseases report) Figure 54 (from section 4.1.1), which provides me with hpv prevalence c.2010 in UK. This is before vaccination really makes a difference to prevalence (esp as the age brackets start at “<25” here), so should approximately work for an initial state.
      * Note that <https://digital.nhs.uk/data-and-information/publications/statistical/cervical-screening-annual/england-2023-24/section-2-cervical-cytology#adequate-samples-gp-and-nhs-community-clinics-> tells us that over 95% of screened women have normal cervical cytology, meaning that although the hpv prevalence data I am using is for women with normal cervical cytology, because that is such a majority of the population, it still works just fine!
      * Big qu here is I am assuming the initial prevalence in men is same as that in women, while the HPVsim model by default puts it somewhat higher in women than in men for the lower age ranges. Still, will only change tis if needed I think
      * Using <https://hpvcentre.net/statistics/reports/GBR.pdf?t=1757520103139>, figure 60, to estimate initial hpv type distribution. The HPVsim model needs this as a probabilitiy distribution (but it will rescale it itself), and any agent that is assigned to start the simulation with HPV will be assigned only one of the genotypes to start with – this is one of the things we need the burnin period for, to get a proper spread of genotypes!
  + rel\_init\_prev is a scale factor for the initial hpv prevalence (set by default to 1, but we could do a calibration which includes this to see if we should scale population-wide hpv prevalence up or down.
    - I could perhaps calibrate this, but only if things really aren’t looking great. I think stuff is fine
  + They are used in *sim.init\_states*
    - There we can see that if init\_hpv\_dist is None (as it is by default), we assume an even split between HPV genotypes which definitely seems wrong.
  + I further will do calibration of some genotype data, which is the genotype distribution of cancers – this data is provided by kate (and I am just mentioning it here as although it doesn’t alter our initial states ofc, it is still about the idea that our hpv prevalence across genotypes in our model needs to reflect reality)
    - This is d3/mesherHPVCancerDist.csv, from <https://doi.org/10.1136/jclinpath-2014-202681>
    - Importantly, the source for this data notes the overlap between HPV16 and HPV18 prevalence; they are not mutually exclusive as 68% have hpv 16, 19% have hpv 18, and 83% have both. However, the lack of overlap is sorta small so for now not bothering and assuming you can only have 1 or the other when you have cancer for calibrations’ sake, but presumably will need to fix this in due course? Or at least could try to.
* *condoms, eff\_condoms* <-for a sensitivity analysis!
  + This parameter gives the proportion of acts in which condoms are used. It is a dictionary indexed by layer type, so we can specify different proportions for the marriage and casual layers
  + Condom use in casual relationships appears to be between 20-60%, but this is quite hard to ascertain and really requires a sensivitiy analysis across 5-40% at 5% intervals perhaps. For marriage seems between 2% and 20%. I will go by that slightly old (20years old ish) BMJ article, and go for 50% condom use in casual relationships, 17% use in marriage
    - *“Turning to partnership characteristics, formation was associated with higher rates of condom use at last sex, with 54.8% of men reporting condom use on this occasion by contrast with only 19.4% of men in partnerships of at least 5years’ duration, while the corresponding percentages for women are 46.3% and 15.7%. Rates of condom use at last sex appear to fall markedly and quickly reach aplateau from 6 months”*
      * *10.1136/sti.2005.019117*
    - [*Contraceptive Use by Method 2019 | Population Division*](https://www.un.org/development/desa/pd/content/contraceptive-use-method-2019)
    - [*https://doi.org/10.1371/journal.pone.0304952*](https://doi.org/10.1371/journal.pone.0304952)
    - [*Alarming decline in adolescent condom use, increased risk of sexually transmitted infections and unintended pregnancies, reveals new WHO report*](https://www.who.int/europe/news-room/29-08-2024-alarming-decline-in-adolescent-condom-use--increased-risk-of-sexually-transmitted-infections-and-unintended-pregnancies--reveals-new-who-report)
  + This is quite different to what it was before, as it was **(m=0.01, c=0.27)**
  + I am keeping *eff\_condoms* at 50% as this is consistent with literature (the first link below is some further literature which aggrees with the second link below, and the second link below is what HPVsim uses to justify setting it to 0.5 by default).
    - [*https://doi.org/10.1093/infdis/jit191*](https://doi.org/10.1093/infdis/jit191)
    - ***10.1056/NEJMoa053284***
* *n\_clusters, cluster\_rel\_sizes, add\_mixing*
  + These all relate to geospatial clusters and mixing between the geospatial clusters. We aren’t adding this to our modelling, so they remain at their default (1 cluster and None for both the other parameters)
* *debut*
  + Age of sexual debut. Defined as two probability distributions, one for females and one for males, which by default is: f~N(15,2.1), m~N(17.6,1.8)
  + I am doing debut=dict(f=dict(dist='normal', par1=16.0, par2=3.1), m=dict(dist='normal', par1=16.0, par2=4.1))
    - I don’t really have much evidence for this. Some data I am looking at seems to say that in the UK most people start having sex at 16 with around a third of both boys and girls starting before 16. But I cant really get much data here so I think this is all for now. Ideally we can do better though.
* *f\_cross\_layer, m\_cross\_layer*<- needs calibration + sensitivity analysis of calibrated values
  + The proportion of females/males who have concurrent cross-layer relationships
  + I think that this only makes sense if it is a value **between 0 and 1 – and looking at its use in the HPVsim source code, it does look like it is used to inform binomial probabilities (see population.py), so I will constrain its range in calibration to this**.
  + These are parameters for calibration – because there is very little data to suggest what they should be. From searching online, it looks like possibly around 25% of men and 15% of women cheat, so although relationships may well be consentually non-monogamous and this is a modelling oversimplicaition, I will use these values as the ‘best guesses’ to supply to Optuna
    - <https://www.techopedia.com/statistics/cheating-statistics>
* *beta* <- needs calibration + sensitivity analysis of calibrated values
  + Per-act transmission probability – this has a default value of 0.25, found through calibration of HPVsim
  + It cant be too crazy small, nor crazy large, so I am allowing it to be in the range [0.0, 0.50] with a best guess of 0.25 as informed by the HPVsim calibration. **This is different to the range it was before, as it was before [0.0,0.2] with a best guess of 0.05 – although this did get good results which were pretty consistent between calibrations, it is inconsistent with what the HPVsim team had and didn’t even give a chance for the beta to agree with their value. If subsequent calibration are bad though, and especially beta is >0.2 in the bad cals, I may want to change this back to how it was before though!**
* *transf2m, transm2f*
  + Relative transmissibility of receptive->insertive (resp insertive->receptive) partners of HPV during intercourse.
  + Calibrated previously
  + **I will not calibrate this for now**, as I don’t see why this should be any different between the setting where they originally calibrated HPVsim and England, once we have accounted for parameters like beta and condom usage.
* ‘parameters for disease progression’, ‘parameters used to calculate immunity’ as labelled in parameters.py
  + I am not planning to tinker with any of these parameters and just stick to the standard HPVsim parameters. They all concern the natural history of the disease, which should surely(?) be the same in England as wherever HPVsim was calibrated for (Nigeria?). So if i assume they did it correct for where HPVsim was originally calibrated to, no need to tinker here.
  + They do suggest in the HPVsim paper’s supplementary materials that we do a sensitivity analysis with *hpv\_control\_prob* varying due to the debate whether it is 0 or not!
* *genotypes*
  + Picking which genotypes will be modelled.
  + Unlike before, where I did not specify *genotypes* (and then presumably just modelled 16 and 18?), I am now modelling all of the genotypes included in the 9-valent vaccine, that is 16,18, hi5 (the 5 other ones included in 9-valent),
    - I am NOT modelling ohr (other high risk types not included in 9-valent), lr (low risk) – perhaps I should do that for a more complete simulation? Or if HPV-PRIME requires it then I will need to tack this on
* Genotype parameters:
  + 16,18
    - *dur\_precin* is based on literature, don’t touch
    - *cin\_fn\_k* is currently calibrating (0.5,0,1) but there are two other pars in that dist too, should they be calibrated?
    - *dur\_cin\_par1* is currently calibrating (6,1,12) but in this lognormal there is a second parameter that is not being calibrated (though this is the variance so perhaps we are deliberately keeping as is)
    - *cancer\_fn* has a transformation prob that is derived from a HPVsim calibration so we will keep as is, and trust their calibration as that should all be the same
    - *rel\_beta\_ and sero\_prob* are chosen according to literature, so don’t touch
  + hi5
    - *dur\_precin* seems to be matching literature according to Supplementary Materials Table S2
    - *cin\_fn\_k* perhaps should be calibrated to match what we do with 16,18
    - *dur\_cin\_par1* perhaps should also be calibrated to match what we do with 16,18
    - *cancer\_fn* has a transformation prob that is derived from a HPVsim calibration so we will keep as is, and trust their calibration as that should all be the same
    - *rel\_beta*  is marked as a PLACEHOLDER so perhaps worth calibrating
    - *sero\_prob* is marked as a placeholder in the code but in Supplementary materials Table S2 it looks like it is informed by literature, so lets keep it at 0.6
  + **^for the above, I think all the yellow stuff may well be calibrated – however also I am justified in not calibrating some or any of these, as they all are from calibrations by the HPVsim team; so although they aren’t from literature (and therefore aren’t gospel and I can go and recalibrate them if I want), it may be a good idea to try a calibration with all the genotype calibration as it was (other than hi5 being like 16 and 18), and then one with all the yellows above being calibrated and see which gets a better result, and going from there to see which things I should further tweak to get the best fit possible to UK data**
* *mixing*
  + sexual mixing matrices by layer type (i.e. one for casual and one for marriage) – this has massively improved calibration quality so clearly was key, and I am using Sophie’s data here. However, I am not sure where her data came from!

Still to work out:

f\_partners, m\_partners, acts, age\_act\_pars, layer\_probs (c. 50% of the uk population is married if that helps! Hpvsim people agree with me that for casual relationships its just assumptions – though looking at how layer\_probs is made in HPVsim it seems pretty much to be going for 50% of people wanting to be married at some point in their life if unpartnered), dur\_pship,

* it says in the supp materials that the following’s default values are ‘assumptions only’ so deffo need to have some informed values or be calibrated (or at least, ideally): dur\_pship, m\_partners, f\_partners, acts (it notes that “very little data is available” – it may be a good idea to simply try to calibrate these negative binomial distributions for coital acts, picking the mean and dispersion parameters for calibration )
  + - Therefore, **for network parameters, if I can’t do an informed estimate of their values using literature, I will stick with the HPVsim default but do a pretty involved sensitivity analysis to show how results vary/stay the same over a large range of possible behaviour. Conversely, for natural history/transmission/other not-sociological parmaeters (i.e. ones where I am just trying to find a ground truth which should remain fixed), my sensitivyt analysis is conceptually different, as it looks to see what if my calibration was off by a bit, by doing jittering/trying out a few of my best cals, etc.**

for dur\_pship, can use <https://marriagefoundation.org.uk/wp-content/uploads/2019/12/MF-note-Average-length-of-marriage.pdf>, which deffo seems to imply that things are different to how it is modelled here and I could at least tweak the negative binomial somehow

* Seems sorta tricky to get the neg binom distrib to work but i think i approximately have it here: <https://www.desmos.com/calculator/epaqumoa8h>, so just need to now work out what parameters are best for both marriages and casual relationships i suppose

I further need to look at every parameter in the parameters.py file and check its got a reasomable value guided by literature – where I want a different value to whatever is default by HPVsim, do that and then that is to be mentioned in my paper, or ofc it is calibrated!

^ this is almost done, just to do the stuff I have noted above about networks and then time to plan a whole lot of lil calibration experiments!

**d2Cal\_14Sep25\_XPS\_E0\_1, d2Cal\_14Sep25\_ZB\_E0\_2**

* Starting off with some simple E calibrations – these are not using any updated network parameters, just calibrating the parmaeters we did before in D with the same ranges as D (only exception being the additional of hi5, which is having its same parameters calibrated as 16 and 18), and the big difference from D is that we are using better informed other parameters:
  + Hi5 included in genotypes
  + Condoms: m=0.17, c=0.5,
  + init\_hpv\_prev and init\_hpv\_dist updated to better match uk
* This may not be the best, and could benefit from:
  + Changed ranges (see results from cals to see if they press against ranges)
  + Updated network parmeters (see above)
  + Calibrating more of the genotype parameters (see above)
  + Calibrating against the D3 data (at least cancerous genotype dist)