**Calibration ...**

We calibrated LungFlag on bi-monthly samples on limited years ('simulation input'). In this report we check the calibration on THIN, Optum and Geisinger.

עדכון מאי 2024

מצגת מתוקנת עם calibration index ובלי "הביצועים הכלליים" נשלחה לליאת לטובת המסמכולוגיה

אימנו את LungFlag על KP, כיילנו את המודל על בסיס דגימות דו-חודשיות מ- THIN משנים מוגבלות.

המסמך סוקר את איכות הכיול ממספר זוויות, ביחס לדגימות אקראיות ב- THIN בכל השנים, Optum ו- Geisinger, תוך הסתכלות גם על PLCO במקרים השונים.

דרך אחת פשטנית להסתכל על הדברים היא לבחון את הביצועים הכלליים - מספר ה- cases שהמודל חוזה בהשוואה למספר שיש ב- cohort. החוט ש(אולי)מקשר בין הדברים הוא יחס די קבוע בין "הביצועים הכלליים", של LungFlag ו- PLCO.

ב- THIN אנחנו רואים שלושה דברים:

* לא במפתיע, LungFlag מכוייל היטב
* PLCO לא מכוייל הערכת יתר של הסיכון עד כדי פקטור 2 ב"ביצועים הכלליים"
* ה- incidence rate ב- cohort נמוך מהספרות

ואז כשעוברים ל- Optum:

* ה- incidence rate דומה לספרות
* PLCO קרוב למכוייל
* והתוצאה, LungFlag רחוק מלהיות מכוייל עם הערכת חסר של הסיכון

התמונה המלאה יותר מסובכת:

* מידת הכיול או האי כיול תלויה בשאלה האם מסתכלים בציונים נמוכים (מרבית האוכלוסייה) או באלו עם סיכון גבוה יותר, יש גם הבדל בין אוכלוסיות (USPSTF, מעשנים וכלל האוכלוסייה).
* מסמך זה משווה PLCO בגילאי 55-75 (רק שם הוא מכוייל) ל- LungFlag בכל הגילאים (ויש גם ממצאים ביחס לכיול של LungFlag בקבוצות גיל שונות, אבל נראה שהם לא בגרעין הדיון)

שאלה רלבנטית נוספת היא האם הניתוח הזה אינו over fitting שאין לנו דרך לבדוק? לכן לקחתי את Geisinger כסוג של validation לתובנות, והתמונה רחוקה מלהיות חלקה:

* היחס בביצועים הכלליים בין המודלים נשמר, אבל, להבדיל מ- THIN ו- Optum, היחס תקף רק/בעיקר בציונים נמוכים, ולבעלי סיכון גבוה הכיול של שני המודלים כמעט מתלכד.
* לגבי incidence rate, נראה שכרגע יש בעייה בחישוב.

איך ממשיכים מכאן?

כיוון שיש עדויות רבות ש- PLCO מכוייל, ועקב זאת שנראה שב- THIN חסרים cases, ולאור שאר הדברים, ולמרות התמונה המורכבת, הנטייה היא לחשוב שנכון לתקן את הכיול שלנו. הדרך הפשוטה ביותר היא פקטור ש"יצמיד" אותנו טוב יותר לכיול של Optum ביחס ל"ביצועים הכלליים", אם כי ייתכן שבציונים גבוהים למשל הוא יפגע בכיול.

בכל מקרה, אנחנו יודעים שהכיול לא ישפיע על הביצועים AUC וכו אבל הוא חשו https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4251899/#pmed.1001764-Tammemgi1ב לפחות לשותפים למאמר.

כדי להחליט מה לעשות יש לענות על מספר שאלות:

* הצורך: מה חשוב בכיול? מה המינימום שאנחנו צריכים להשיג ולהציג במאמר? האם די ב"ביצועים הכלליים" או שנדרש פירוט רב יותר?
* האילוצים: כמה זמן יש לנו?
* המסגרת: עם מי אפשר להתייעץ? כמה פתוחים ומתי אנחנו רוצים להיות עם "השותפים" למחקר

כדי להוסיף לתחושת אי הנוחות אציין ש- Baldwin מצא את PLCO לא מכוייל על Qresearch שהוא דטהבייס בריטי, אבל בכיוונים הפוכים לאלו שאני חושב שאני רואה ... ראה עמוד 20 ב- https://www.thelancet.com/cms/10.1016/S2213-2600(23)00050-4/attachment/b0edbf5d-20f2-47cb-9722-72097349b9fb/mmc2.pdf

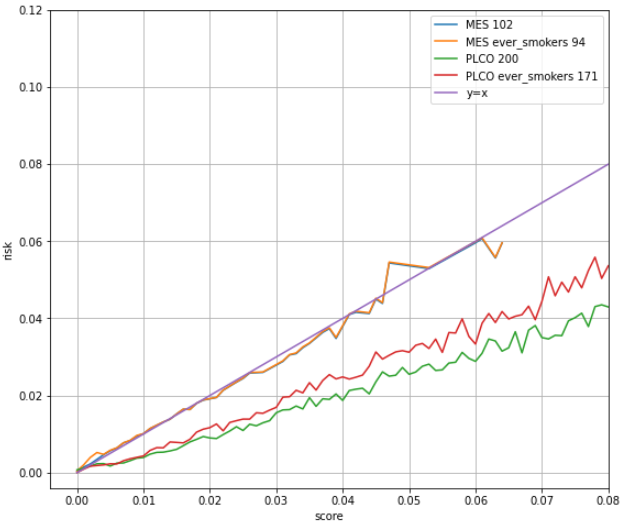
**THIN – Validate Calibration**

Graph T1 checks the calibration with ‘2 samples per year, random time’, time window 730 (0-730, suspected are not excluded), **for ‘ever smokers’ and ‘all’ populations**.

The legend includes ‘average performance’ in %:

‘average performance’ = 100 \* expected number of cases / actual number of cases

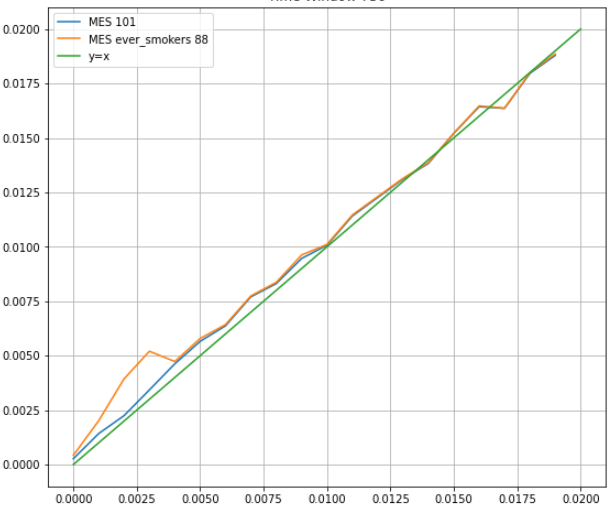
Thus, for instance, ‘average performance’ of 150% means the model predicts 50% more cases than the actual number of cases.



Graph T1 – Calibration Check on THIN, time window 730

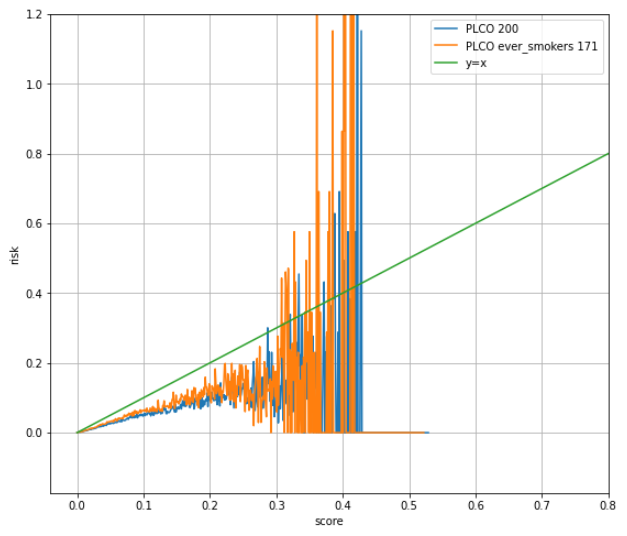
What can be seen:

* X-Axis: Score, Y-Axis: Risk, so line y=x is perfect calibration.
* LungFlag:
  + As expected, LungFlag is calibrated:
  + Average performance is 102% for the full population and 94% for ever smokers.
  + Note that there is for most of the graph there is no visible difference between ‘ever smokers’ and ‘all’ populations, but there is a significant difference in the left end (very low scores) where for ‘ever smokers’ the risk is higher than predicted – see zoom next. We think it might be patients with score biased by ‘wrong’ imputation.



Graph T1a – Zoom left corner (just LungFlag)

* PLCO:
  + PLCO score is calibrated to 6 years. As we investigate time window 730:
    - The naive (and wrong) approach would be to divide the score by 3.
    - Why wrong? Because risk grows with age. From incidence rate in literature in the age range with most patients, the rate is ~7% per year.
    - As a result, in the first 2 years just 1 / 3.46 of the ‘cases by 6 years’ are expected and used in the graph.
  + PLCO score is calibrated just for the age range 55-75, so we drop older and younger patients.
  + PLCO2012 is used for ‘ever smokers’. The risk in the cohort is much lower than projected. Average performance 171%.
  + PLCO2014 includes ‘never smokers’, and is used for ‘all’ population. The graph shows It is underestimating the risk in the cohort as well. Average performance 200%.
  + PLCO score can be higher than presented in T1, so next we present PLCO calibration for higher score. The trend is the same, though we have more noise due to small numbers. Note that PLCO bar is 1.5%, equivalent here to score ~0.052 (1.5 \*3.456).



Graph T1b – Extended score range, just PLCO

Implications?

A possible reason for PLCO calibration error might be the factor we calculated – how to move from 6 years risk to 2 years risk. However, it is not likely, as it seems right when looking at other datasets.

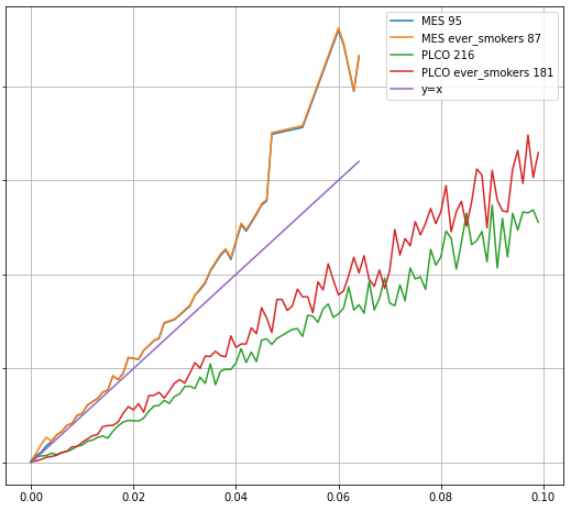
Another possible reason is bias in the representation of cases and control in the cohort - if cases were sub-samples by factor ~2, then PLCO would have been calibrated.

In support to this option, note that we show later significant differences (though not factor 2) between cohort incidence rate and actual incidence rate.

The implication is that we might have calibrated LungFlag on case-control biased dataset.

**Time Window 365**

We repeated the analysis with time window 365 – see graph T2. What do we see?

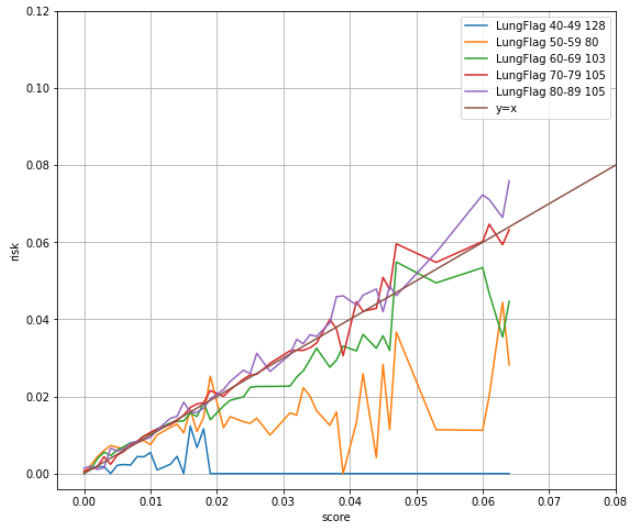


Graph T2 – Calibration Check on THIN, time window 365

LungFlag is less calibrated (compared to time window 730) mostly at high scores.   
This is probably because for LungFlag flagged patients, the risk in the first year is higher than the risk in the second year.   
Average performance is just 95%, and 87% for ‘ever smokers’.  
  
PLCO results are similar to time window 730.

**THIN – Calibration by Age Groups**

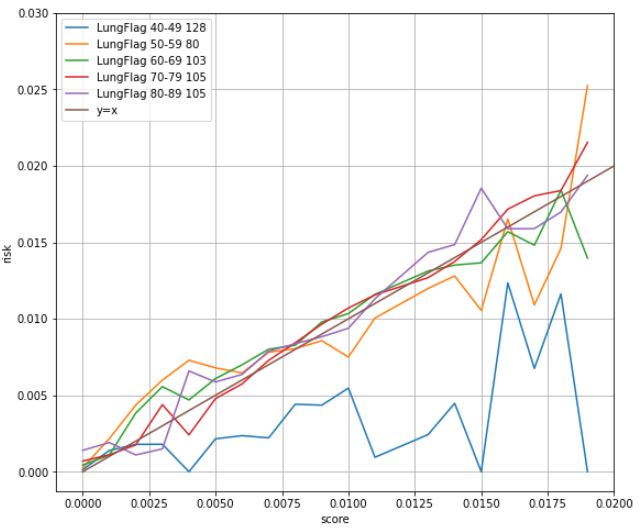
The following graphs present the calibration, for LungFlag, ‘all’ population, time window 730, by age group.



Graph T3– Calibration Check on THIN, different age groups

Most of the cases are age 70 and above, and with the relevant age groups the calibration is almost perfect, as expected. However, as we look at younger and younger age groups, and high scores, the ‘calibrated’ score overestimates the risk.

At the same time, when we zoom on low scores (where most of the scores are), then we see that we underestimate the risk for age group 50-59. It might be interesting but also cohort specific, so before coming to any conclusion - we need to check whether it is the case with other datasets as well.



Graph T3 – Zoom left corner

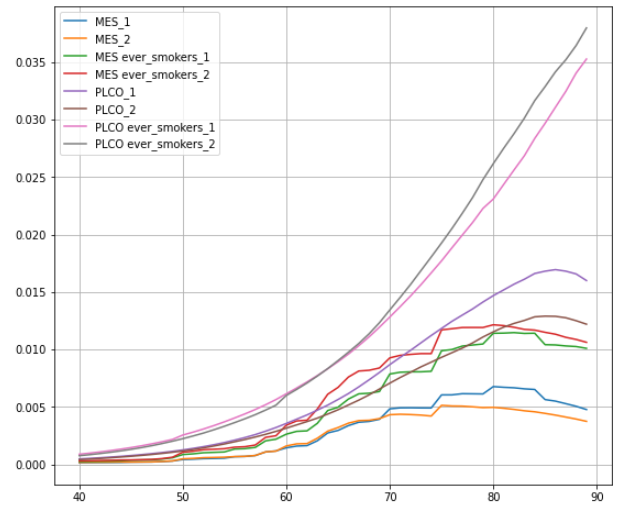
**THIN – Risk by Age**

Graph T4 plot (Y-Axis) average score (estimated risk) per:

* Age: X-Axis.
* Gender: 1 – Male, 2 – Female.
* Population: all (== LungFlag & PLCO2014) or ‘ever smokers’ (== ‘plco2012’).

What can be seen:

* Overall, the behavior is as expected.
* As expected from checking the score calibration, PLCO 'risk per age’ is higher than LungFlag.
* As it is known (and see next), risk by age is higher for Male than Female.
* LungFlag ‘risk per age’ graph is staircase:
  + Reason not clear.
  + In real-life we expect a smoother curve.
  + Not seems important though.



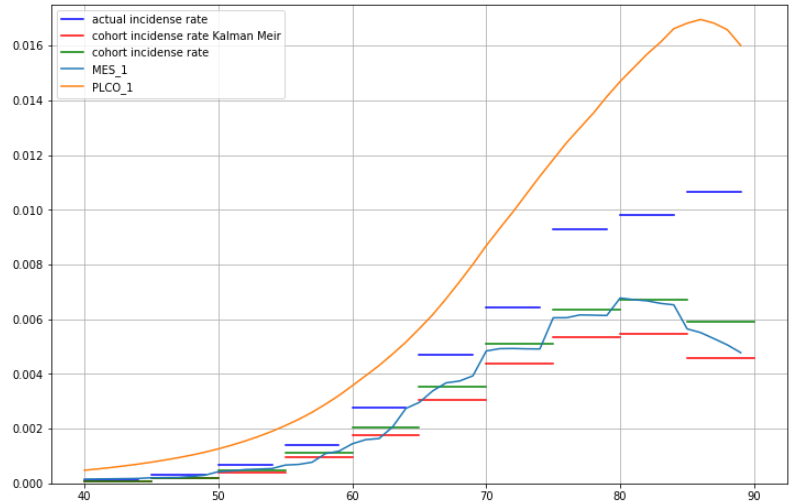
Graph T4 – Risk by Age

**Is THIN representative?**

When we calibrate on THIN and test on other cohorts, we wish THIN is ‘good’ representation of reality (though it is not clear how important it is).

We have a program in Flow for cohort incidence rate calculation with two options – with or without Kalman Meir. ‘With’ is more accurate.

For reference, we downloaded incidence rate (2017 edition) from <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancerregistrationstatisticscancerregistrationstatisticsengland>.



Graph T5 – Incidence rate

What can be seen:

* X-axis: Age, Y-axis: Risk per 2 years
* Quite significant differences between the incidence rate in the reference (blue) and the cohort (green / red). See previous discussion about possible implications.
* The risk per age the model predicts is close to the incidence rate in the samples. However, LungFlag risk (light blue) is closer to the incidence rate calculated without Kalman Meir, might be related to the way we sample.
* As expected from previous results (plco2014 predicted risk is too high).
* The graph is just for Male. We plotted similar graph for Female – same results.

We checked also age distribution, comparing the cohort to UK statistics from <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2017>. Here the fitness is very good.

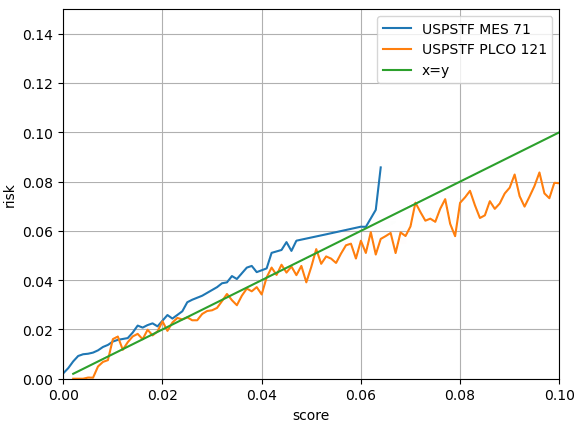
**Optum – Check Calibration – USPSTF**

We checked LungFlag and PLCO calibration on different cohorts – USPSTF, Ever Smokers (for PLCO, limited to age 55-75), All population (for PLCO, limited to age 55-75)

The samples are monthly, 2012-2017 (same samples used for the simulation). Graphs are as before: X-axis: score, Y-axis: risk.

Calibration assessment considered controls sub-sampling by factor 20.

We use PLCO2014 even for USPSTF and Ever Smokers (simpler, however previously we saw that PLCO2014 performances for these group is quite like PLCO2012, so it might have been better to check PLCO2012 calibration directly. Next time ...).



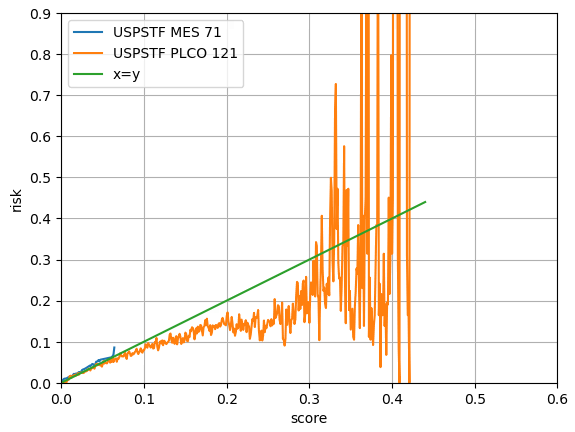
Graph O6 – Optum USPSTF Calibration

USPSTF / LungFlag calibration:

* In first glance, calibration looks quite good at high scores/risk. However, overall we have significant underestimation. The number of cases we predict is just 71% of the actual number. Most of the ‘missing cases’ are accumulated over the very low scores (that are most common).
  + Note that we can’t suggest it is an imputation issue, as USPSTF patients have full smoking info.
  + It might be related to the issues with THIN case-control population used for the calibration, as discussed before.
* Note also the spike in the highest score, probably limitation of our ability to collect enough patients in the highest risk bin in THIN.

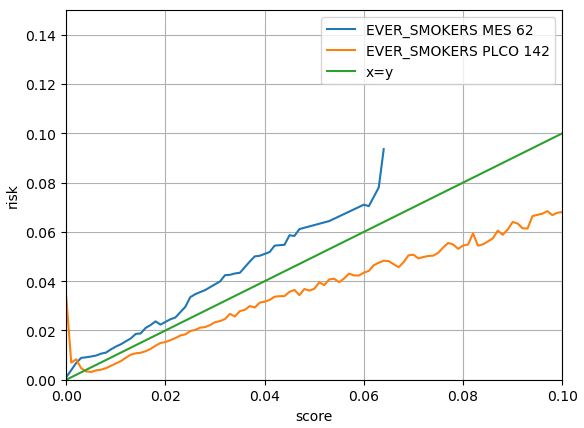
USPSTF / PLCO calibration:

USPSTF / PLCO calibration is almost perfect, until you reach high scores, where PLCO overestimates the risk. The next graph extends the range of scores in O6 – the trend continues even though the results have more noise due to small numbers.



Graph O6b – Extended score range

**Optum – Check Calibration – Ever Smokers**



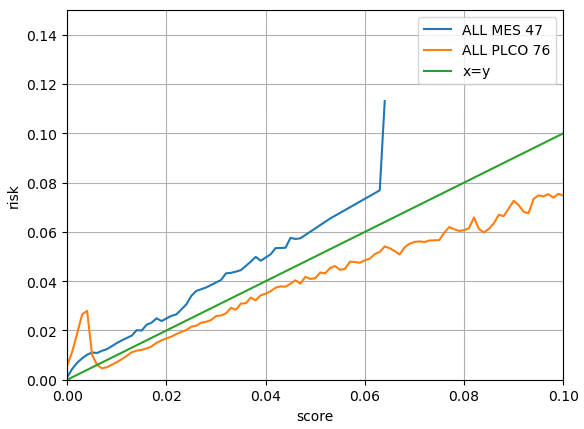
Graph O7 - Optum Ever Smokers Calibration

LungFlag (again) underestimated the risk, even more than within USPSTF (average performance 62% compared to 71%). Similar spike at the highest possible score.

PLCO results deteriorated, and the model significantly overestimated the risk. Could be due to smoking features imputation limitations.

What might be most important, is the ratio between the average performance of PLCO and LungFlag, which is like the ratio in THIN. If we assume that PLCO is calibrated with real balanced data, then we might ‘carry’ a bias because we calibrated LungFlag on (might be biased ...) THIN. See next similar check with Geisinger.

**Optum – Check Calibration – All population**



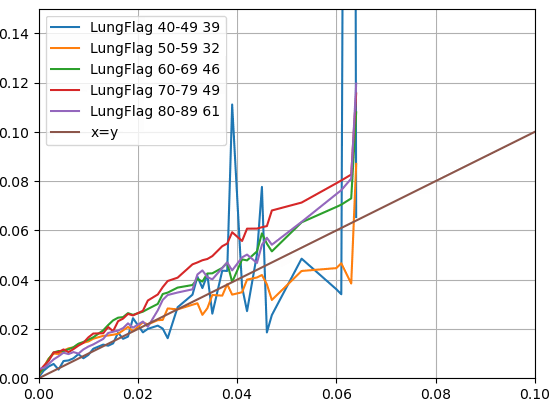
Graph O8 - Optum 'all population’ Calibration

LungFlag (again ...) underestimated the risk, even more than USPSTF or ever\_smokers population. Also, Similar spike at the highest possible score. Note that the overestimation is relatively stronger at low scores, that are most common.

PLCO:

* For high scores, PLCO results overestimate the risk as before.
* However, for low scores (which are common), it is strongly underestimating the risk, as expected, as PLCO is not calibrated for never smokers.
* So overall, the average performance of PLCO is 76% compared to 121-142% before.

**Optum – LungFlag Calibration by Age Group**



Graph O9 – Optum LungFlag Calibration by Age Groups

Most of the cases are in older age groups, so the calibration we see for age group 70-80 for instance, is the same as the calibration we saw in graph O8.

For younger age groups, we see two trends:

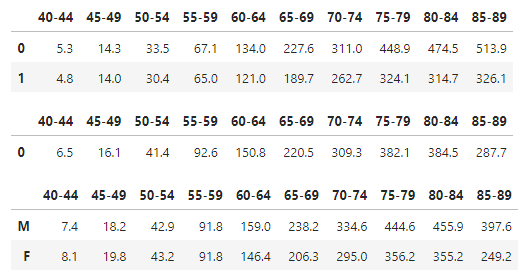
* Overall, stronger underestimation of the risk, dominated by the low scores
* For high scores, the risk is somewhat overestimated

In any case, we don’t see a specific concern with low scores risk underestimation for age group 50-59, as we saw in THIN.

**Optum – Risk per Age and Incidence Rate**

We found 2 sources for actual incidence rate, both for 2016-2020 (As the cohort is 2012-2017, numbers should be a bit higher, but we ignore it):

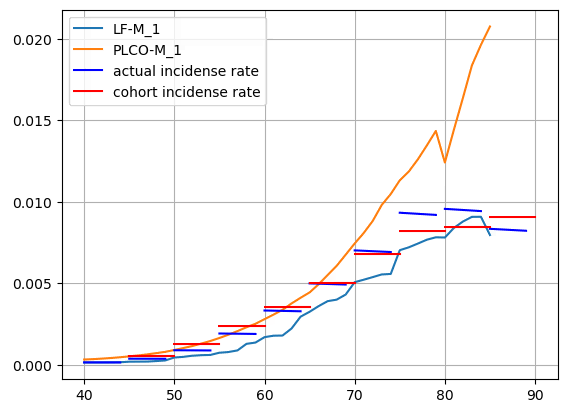
* <https://gis.cdc.gov/Cancer/USCS/#/Demographics/> without separation Male/Female
* <https://www.health.ny.gov/statistics/cancer/registry/table6/tb6lungnys.htm> for the state of NY
* UK numbers are presented just for reference.

   
UK Male  
   
UK Female  
  
  
US  
  
  
NY Male  
   
NY Female

In the US first data set numbers are, for most of the age groups, between Male numbers and Female numbers for NY. Therefore, we use NY numbers.

In the following graph:

* In literature we have incidence rate for 1 year. To get 2 years, we multiplied by 2.07, see previous discussion.
* Cohort incidence rate was calculated using Kalman Meir – see previous discussion.



Graph O10 – Optum Incidence Rate – Males

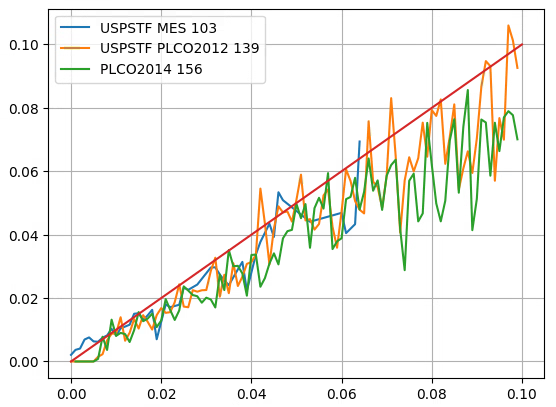
What can be seen:

* The incidence rate in the cohort is very similar to the literature.
* LungFlag underestimates the risk.
* PLCO overestimates the risk and is far off for age > 75 (and also for age < 55, but that is hard to see in the graph).

We checked and got similar results for females.

**Geisinger – second validation dataset**

**USPSTF**



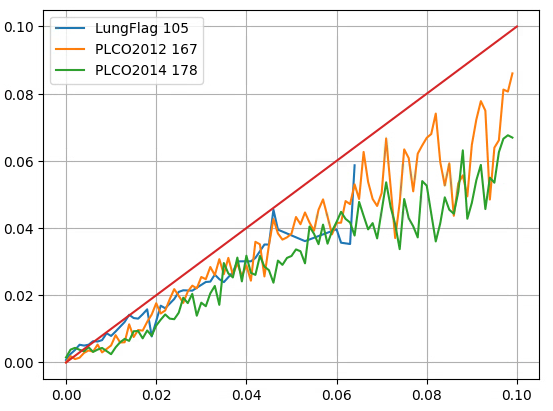
Graph G11 –Geisinger USPSTF calibration

Average performance of LungFlag is almost perfect, though we have risk underestimation at low score and risk overestimation at high scores.

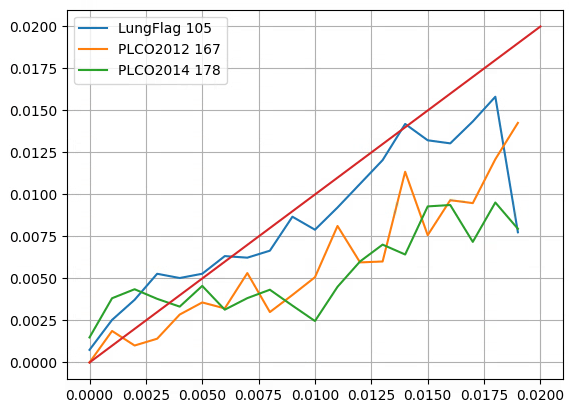
PLCO is not as calibrated as we saw in Optum.

What might be most important is that the ratio in average performance of LungFlag and PLCO2014 (that was used in Optum) is like the ratio between the two that we saw in Optum. See next similar conclusion for ever smokers.

**Ever Smokers**



Graph G12 – Geisinger Ever Smokers calibration



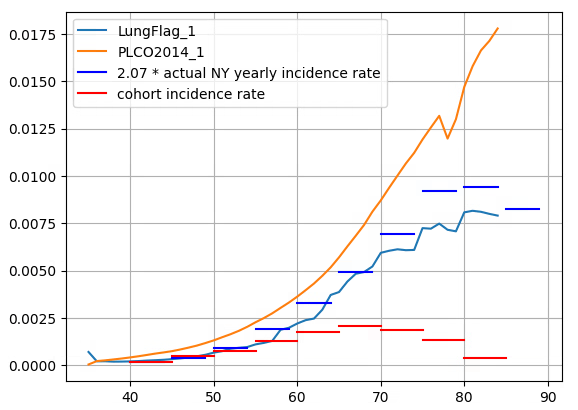
Graph G12a – Zoom on low scores

First recall that for PLCO age range is 55-75, while for LungFlag it is 40-89.

The results are like with USPSTF, see discussion above. Note that for Ever Smokers the major difference between LungFlag calibration and PLCO is at low scores (which are most common), as can be seen in G12a.

**Geisinger – Risk per Age and Incidence Rate probably a bug**

The following graph shows the incidence rate in Geisinger compared to predicted, and actual incidence rate in Ney York. The graph does not make sense ...



**Graph G13 – Geisinger risk by age**

Why does it looks like a bug? See next the calibration per age group. For 70-80 for instance, the average performances in the graph below are good, not like in the graph above.

