Leveraging Advanced Analytics to enhance vaccination outreach

Humana-Mays 2021 case competition

Humana

TEXAS A&M UNIVERSITY
Mays Business School

Table of Contents

1 Executive Summary

In the US alone **there have been 42Mn cases of COVID-19 and 674k deaths** since the outbreak of the pandemic

Vaccines have been established as the best mitigation mechanism to reduce the risks of the virus – 10-fold effectiveness for symptom reduction and 29-fold reduction in likelihood of hospitalization[1]

~45%[2] of the adult US population (147Mn) is still unvaccinated – the most prevalent reasons being side effect concerns, skepticism of its effectiveness and disbelief of its need

Having a sizable share of the population unvaccinated carries three negative consequences – **Increased humanitarian health risk, global economic slowdown, and monetary cost disbursements for health insurers**

At an average financial cost of hospitalization of $21k, **studies[3] indicate that the US is paying >$1Bn per month in hospitalizations of unvaccinated Medicare patients alone**

In order to increase the vaccination intake, most efforts[4] focus on three mechanisms **– Increasing vaccination access, boosting vaccination demand, and overcoming practice-related barriers**

Humana has partnered with *Mays Business School* to explore alternatives to boost the vaccination intake through an analytics case competition. **Our suggested solution consists of a 5-step framework with the objective to increase the vaccine demand by quantifying the benefit of potential interventions and finding an optimal assignment**:

1. **Quantify individual cost of risk** - cost for *Humana* in case of hospitalization
2. **Characterize potential interventions** - cost and effectiveness of some possible actions Humana can take
3. **Estimate the hesitancy of vaccination** - degree to which a *Humana* member does not want the vaccine
4. **Optimize intervention assignment** - assignment of actions to members based on the previous metrics
5. **Adjust for fairness** - control final outreach policy to control for hidden biases

Following this framework we have created a cost-effective prioritized list of pairs member-intervention and **we suggest doing a first touchpoint sending a digital reminder to 3% of *Humana* members and providing an educational newsletter to 97% and a second touch point to 62% of *Humana* members with a phone call, 2% with medical discount, 28% with cash disbursement and not intervening for 8%**

We have sized the potential impact for *Humana* in ~19Mn USD in case of roll-out for the unvaccinated members under study – Since data from March has been used and much has changed since then, in order to make the results of this study actionable, the prioritization would need to be re-computed with up-to-date data that reflected the present rates of vaccination

[1] https://www.cnbc.com/2021/08/24/cdc-study-shows-unvaccinated-people-are-29-times-more-likely-to-be-hospitalized-with-covid.html

[2] https://www.washingtonpost.com/nation/2021/09/28/covid-delta-variant-live-updates/

[3] https://www.usnews.com/news/health-news/articles/2021-09-10/average-covid-hospitalization-is-150-times-more-expensive-than-vaccination

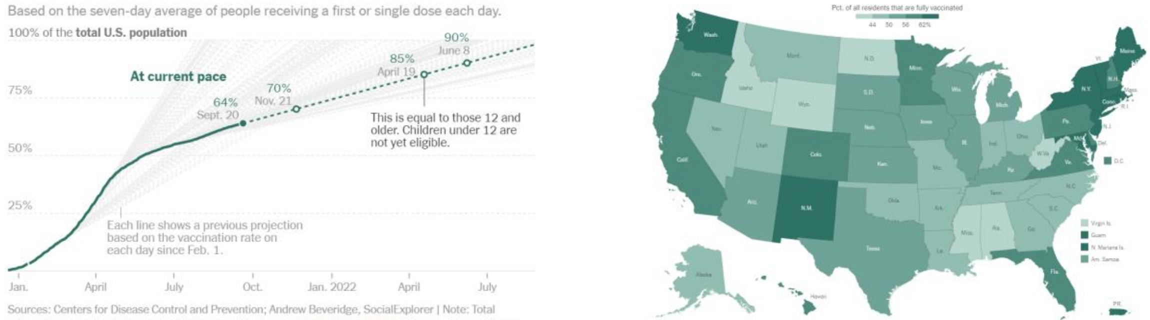
[4] https://www.amjmed.com/article/S0002-9343(08)00466-X/pdf

2 Our approach to solve the vaccination problem

Vaccination is a practice with a long-standing history. Historians date it back to as early as 1000 CE when the Chinese employed smallpox inoculation setting up the basis for the future innovations from Edward Jenner who in 1796 used cowpox to create immunity to smallpox. Despite the scientific and historical evidence that vaccines are the gold-standard to mitigate the risks of diseases, as seen with Smallpox, the Plague, and the Yellow Fever to name a few, they have always suffered from detractors who question their efficacy[5]. Nowadays it is estimated that immunization efforts prevent 2-3 million deaths per year[6] worldwide.

In the context of vaccine roll-out for the recent COVID-19 pandemic, we observe that a significant share of the US population was vaccinated in the early stages, but the pace has slowed down, projections estimate that it will take until mid-2022 to reach 90% vaccination rate among the US population. These differences in vaccine intake are uneven by state as can be observed in exhibit 2-1.

EXHIBIT 2-1[7]

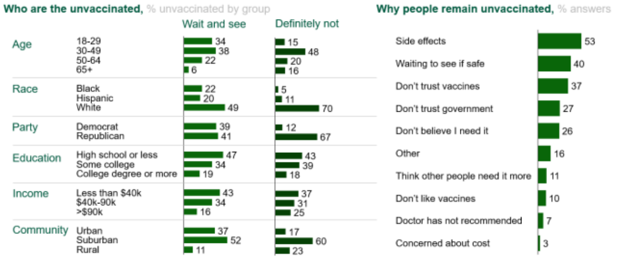


Sources: Centers for Disease Control and Prevention; Andrew Beveridge, SocialExplorer | Note: Total population includes states, territories and three countries with special agreements with the United States: Palau, Micronesia and the Marshall Islands.

Source: Centers for Disease Control and Prevention

Double clicking on the population that remains unvaccinated we can unveil that there are major discrepancies both in terms of socio demographic characteristics as well as causes. White citizens seem to be more averse to vaccination than blacks for instance. In terms of the causes, side effects and trust are the main concerns of the public. Exhibit 2-2 illustrates the discrepancies in vaccine intake both in terms of groups as well as causes.

EXHIBIT 2-2[8]



[5] https://www.historyofvaccines.org/timeline/all

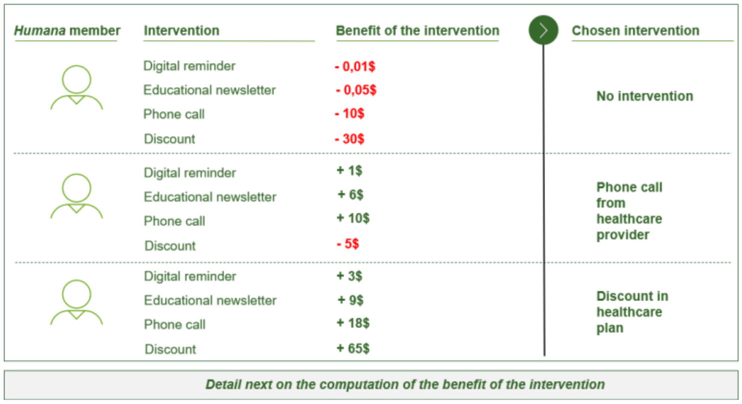
[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4514191/#R3

[7] https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html?auth=login-email&login=email

[8] https://www.nytimes.com/2021/07/31/us/virus-unvaccinated-americans.html

Having the historical frame, context of the existing vaccination problematic and implications for a leading health insurer in mind, we proceed to explain our approach to fast-track the targeted vaccination outreach. In order to calculate the most cost-effective vaccination outreach policy, we have defined a set of potential interventions and calculated the individual benefit of each intervention for every *Humana* member. Our goal is to identify the pairs of intervention-member that yield the highest benefit for *Humana* as seen in exhibit 2-3.

EXHIBIT 2-3

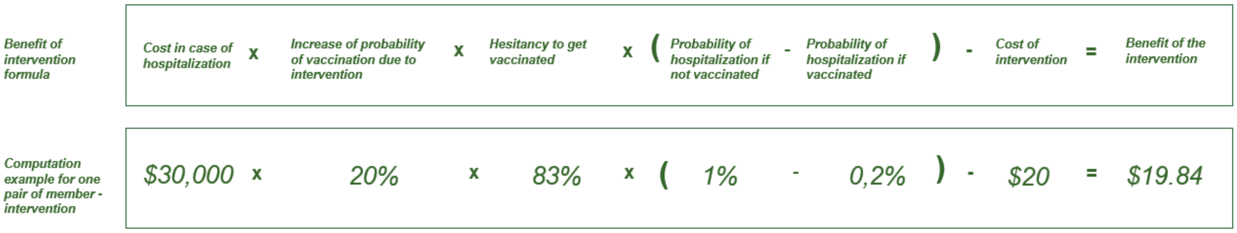


Humana member	Intervention	Benefit of the intervention	Chosen intervention
	Digital reminder Educational newsletter Phone call Discount	- 0,01$ - 0,05$ - 10$ - 30$	No intervention
	Digital reminder Educational newsletter Phone call Discount	+ 1$ + 6$ + 10$ - 5$	Phone call from healthcare provider
	Digital reminder Educational newsletter Phone call Discount	+ 3$ + 9$ + 18$ + 65$	Discount in healthcare plan

Detail next on the computation of the benefit of the intervention

More specifically, for every *Humana* member we compute the benefit of every potential interventions as:

EXHIBIT 2-4



Benefit of intervention formula	*Cost in case of hospitalization*	x	*Increase of probability of vaccination due to intervention*	x	*Hesitancy to get vaccinated*	x	(	*Probability of hospitalization if not vaccinated*	-	*Probability of hospitalization if vaccinated*	)	-	*Cost of intervention*	=	*Benefit of the intervention*
Computation example for one pair of member - intervention	*$30,000*	x	*20%*	x	*83%*	x	(	*1%*	-	*0,2%*	)	-	*$20*	=	*$19.84*

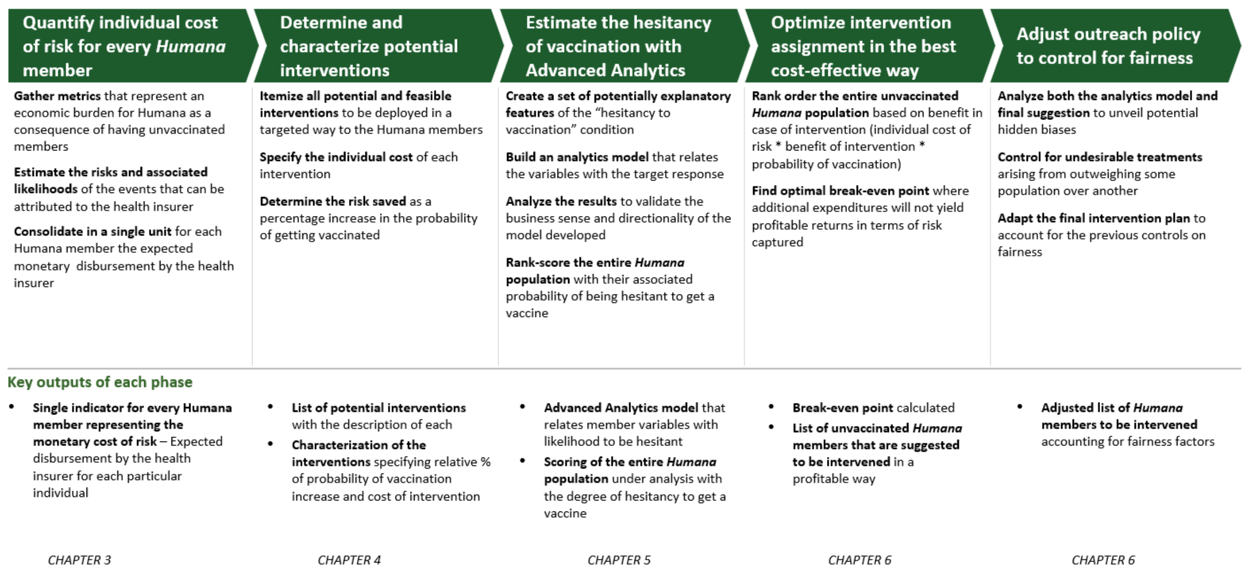
Where the previous terms stand for…

- **Cost in case of hospitalization:** Monetary value in case of hospitalization. It is calculated as an outside-in estimate, for some members like elderly or with existing preconditions it will be higher than for younger and healthier individuals
- **Increase of probability of vaccination due to intervention**: Effect of the intervention in decreasing the hesitancy to get vaccinated. A *Humana* member might respond very positively to a discount while not to an SMS. These values are estimated with an outside-in study where we analyze which interventions have worked best in the past
- **Hesitancy to get vaccinated**: Value which reflects (1 – probability to get vaccinated). It is calculated analytically from the proprietary *Humana* data and the historical member characteristics
- **Probability hospitalization if not vaccinated – probability hospitalization if vaccinated**: Difference in the probability of hospitalization thanks to getting the vaccine
- **Cost of intervention:** Monetary cost of the individual intervention

Having the previous values will allow us to create a grid with all the possible combinations of intervention-member and decide which intervention we would like to assign in a structured and analytical way.

Once we have defined the goal of our approach, that is, compute the benefit of every intervention for every member, we have structured our study as depicted in exhibit 2-5 with a 5-step framework.

EXHIBIT 2-5



Quantify individual cost of risk for every *Humana* member	Determine and characterize potential interventions	Estimate the hesitancy of vaccination with Advanced Analytics	Optimize intervention assignment in the best cost-effective way	Adjust outreach policy to control for fairness
Gather metrics that represent an economic burden for Humana as a consequence of having unvaccinated members **Estimate the risks and associated likelihoods** of the events that can be attributed to the health insurer **Consolidate in a single unit** for each Humana member the expected monetary disbursement by the health insurer	**Itemize all potential and feasible interventions** to be deployed in a targeted way to the Humana members **Specify the individual cost** of each intervention **Determine the risk saved** as a percentage increase in the probability of getting vaccinated	**Create a set of potentially explanatory features** of the "hesitancy to vaccination" condition **Build an analytics model** that relates the variables with the target response **Analyze the results** to validate the business sense and directionality of the model developed **Rank-score the entire *Humana* population** with their associated probability of being hesitant to get a vaccine	**Rank order the entire unvaccinated *Humana* population** based on benefit in case of intervention (individual cost of risk * benefit of intervention * probability of vaccination) **Find optimal break-even point** where additional expenditures will not yield profitable returns in terms of risk captured	**Analyze both the analytics model and final suggestion** to unveil potential hidden biases **Control for undesirable treatments** arising from outweighing some population over another **Adapt the final intervention plan** to account for the previous controls on fairness
Key outputs of each phase				
• **Single indicator for every Humana member representing the monetary cost of risk** – Expected disbursement by the health insurer for each particular individual	• **List of potential interventions** with the description of each • **Characterization of the interventions** specifying relative % of probability of vaccination increase and cost of intervention	• **Advanced Analytics model** that relates member variables with likelihood to be hesitant • **Scoring of the entire *Humana* population** under analysis with the degree of hesitancy to get a vaccine	• **Break-even point** calculated • **List of unvaccinated *Humana* members that are suggested to be intervened** in a profitable way	• **Adjusted list of *Humana* members to be intervened** accounting for fairness factors
CHAPTER 3	*CHAPTER 4*	*CHAPTER 5*	*CHAPTER 6*	*CHAPTER 6*

Following this framework, we will be able to quantify the monetary benefit of every potential intervention to each *Humana* member. The formula in exhibit 4 describes how we reach this final computation.

As noted in exhibit 2-5, the following chapters will cover the main steps sequentially. Chapter 3 is dedicated to studying the cost of risk, chapter 4 to analyzing potential interventions and chapter 5 to develop the probability of vaccination model.

Finally, once these ingredients are in hand, we will proceed with the intervention assignment in chapter 6. Before going any further, we stress some key benefits of the previously defined framework:

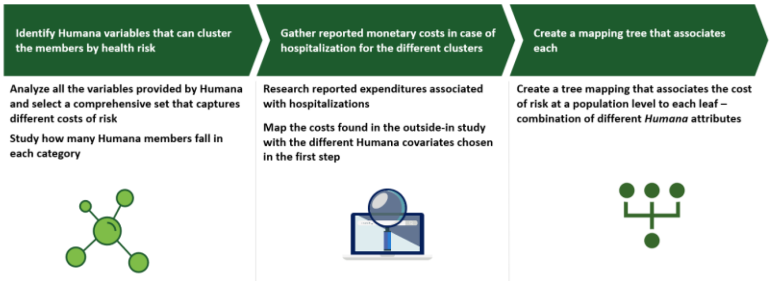
- **Targeted outreach:** This framework provides member-level recommendations
- **Precise quantification of the hesitancy to vaccinate** Thanks to the creation of an analytics model, we are able to quantitatively estimate the hesitancy to vaccinate of every member
- **Accountability for the individual effect of interventions:** The members with highest hesitancy to vaccinate may not be the optimal to be reached out – based on predisposition to embrace the intervention, this framework accounts for the individual effect of interventions at a member level
- **Monetary association of the actions**: Quantifying all the factors in terms of probabilities of certain outcomes and monetary cost allows to operate with them and retrieve a dollar figure for the benefit of each intervention for each *Humana* member
- **Profitability assessment**: Once all the data is collected and the previous indicators created, this framework benefits from calculating the exact breakeven point where it is no longer profitable to intervene – all the decisions are driven by their profitability rather than a qualitative assessment
- **Flexibility to incorporate additional mitigating actions:** If in the future additional potential interventions are devised, they can easily be incorporated to the existing problem and reweight the final decisions
- **Flexibility to change expected costs in case of economic changes:** Similar to the previous one, the framework can be iterated and refined with no additional efforts changing the parameters of the actions

An exhaustive list with assumptions and potential enhancements to this framework can be found in the appendix.

3 Quantifying the cost of risk of not being vaccinated

In this third chapter and throughout the entirety of this study we define cost of risk as the expected monetary disbursement a health insurer might incur in case of hospitalization. Although cost of risk is not an explicit metric in the competition's description, we believe that any attempt to recommend a policy concerning members with heterogeneous conditions and latent risks should account for risk discrepancies. The basic motivation behind the use of such an indicator is that it is not the same to successfully distribute 100 vaccines to healthy and immunologically strong individuals than administering 100 to a group of at-risk individuals. Since there is no tailored-to-*Humana* data regarding cost of hospitalization at our disposal, we have estimated the cost of risk as a function of some key covariates from *Humana* members and an outside-in study of reported costs. Exhibit 3-1 depicts the followed procedure.

EXHIBIT 3-1



Identify Humana variables that can cluster the members by health risk

Analyze all the variables provided by Humana and select a comprehensive set that captures different costs of risk

Study how many Humana members fall in each category

Gather reported monetary costs in case of hospitalization for the different clusters

Research reported expenditures associated with hospitalizations

Map the costs found in the outside-in study with the different Humana covariates chosen in the first step

Create a mapping tree that associates each

Create a tree mapping that associates the cost of risk at a population level to each leaf – combination of different *Humana* attributes

With this approach, we have found external studies[9] and cost breakdowns[10] that are summarized in exhibit 3-2.

EXHIBIT 3-2

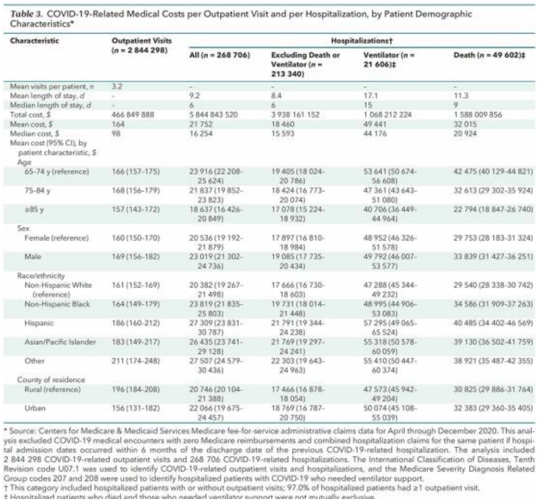


Table 3. COVID-19-Related Medical Costs per Outpatient Visit and per Hospitalization, by Patient Demographic Characteristics*

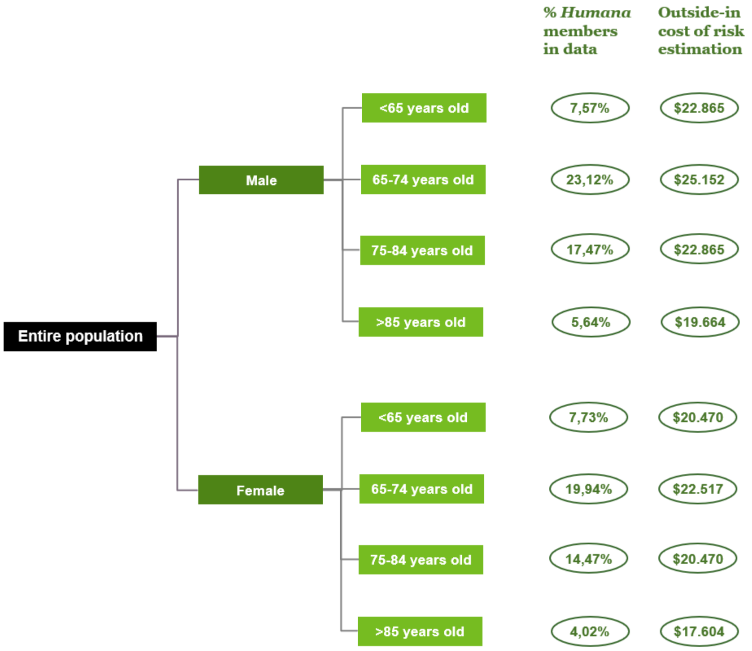
Characteristic	Outpatient Visits (n = 2 844 298)	Hospitalizations†			
		All (n = 268 706)	Excluding Death or Ventilator (n = 213 340)	Ventilator (n = 21 606)‡	Death (n = 49 602)‡
Mean visits per patient, n	3.2	-	-	-	-
Mean length of stay, d	-	9.2	8.4	17.1	11.3
Median length of stay, d	-	6	6	15	9
Total cost, $	466 849 888	5 844 843 520	3 938 161 152	1 068 212 224	1 588 009 856
Mean cost, $	164	21 752	18 460	49 441	32 015
Median cost, $	98	16 254	15 593	44 176	20 924
Mean cost (95% CI), by patient characteristic, $					
Age					
65-74 y (reference)	166 (157-175)	23 916 (22 208-25 624)	19 405 (18 024-20 786)	53 641 (50 674-56 608)	42 475 (40 129-44 821)
75-84 y	168 (156-179)	21 837 (19 852-23 823)	18 424 (16 773-20 074)	47 361 (43 643-51 080)	32 613 (29 302-35 924)
≥85 y	157 (143-172)	18 637 (16 426-20 849)	17 078 (15 224-18 932)	40 706 (36 449-44 964)	22 794 (18 847-26 740)
Sex					
Female (reference)	160 (150-170)	20 536 (19 192-21 879)	17 897 (16 810-18 984)	48 952 (46 326-51 578)	29 753 (28 183-31 324)
Male	169 (156-182)	23 019 (21 302-24 736)	19 085 (17 735-20 434)	49 792 (46 007-53 577)	33 839 (31 427-36 251)
Race/ethnicity					
Non-Hispanic White (reference)	161 (152-169)	20 382 (19 267-21 498)	17 666 (16 730-18 603)	47 288 (45 344-49 232)	29 540 (28 338-30 742)
Non-Hispanic Black	164 (149-179)	23 819 (21 835-25 803)	19 731 (18 014-21 448)	48 995 (44 906-53 083)	34 586 (31 909-37 263)
Hispanic	186 (160-212)	27 309 (23 831-30 787)	21 791 (19 344-24 238)	57 295 (49 065-65 524)	40 485 (34 402-46 569)
Asian/Pacific Islander	183 (149-217)	26 435 (23 741-29 128)	21 769 (19 297-24 241)	55 318 (50 578-60 059)	39 130 (36 502-41 759)
Other	211 (174-248)	27 507 (24 579-30 436)	22 303 (19 643-24 963)	55 410 (50 447-60 374)	38 921 (35 487-42 355)
County of residence					
Rural (reference)	196 (184-208)	20 746 (20 104-21 388)	17 466 (16 878-18 054)	47 573 (45 942-49 204)	30 825 (29 886-31 764)
Urban	156 (131-182)	22 066 (19 675-24 457)	18 769 (16 787-20 750)	50 074 (45 108-55 039)	32 383 (29 360-35 405)

* Source: Centers for Medicare & Medicaid Services Medicare fee-for-service administrative claims data for April through December 2020. This analysis excluded COVID-19 medical encounters with zero Medicare reimbursements and combined hospitalization claims for the same patient if hospital admission dates occurred within 6 months of the discharge date of the previous COVID-19-related hospitalization. The analysis included 2 844 298 COVID-19-related outpatient visits and 268 706 COVID-19-related hospitalizations. The International Classification of Diseases, Tenth Revision code U07.1 was used to identify COVID-19-related outpatient visits and hospitalizations, and the Medicare Severity Diagnosis Related Group codes 207 and 208 were used to identify hospitalized patients with COVID-19 who needed ventilator support.
† This category included hospitalized patients with or without outpatient visits; 97.0% of hospitalized patients had ≥1 outpatient visit.
‡ Hospitalized patients who died and those who needed ventilator support were not mutually exclusive.

[9] https://www.cbsnews.com/news/covid-vaccination-health-insurance-cost-treatment/
[10] https://www.acpjournals.org/doi/10.7326/M21-1102#t3-M211102

Next, we have curated the tree in exhibit 3-3. This tree allows us to group members with similar risk covariates and assign a monetary expected cost of hospitalization. That is, in case of hospitalization of a 70-year-old male, the average cost associated with the hospitalization is of $25.152 for example.

EXHIBIT 3-3



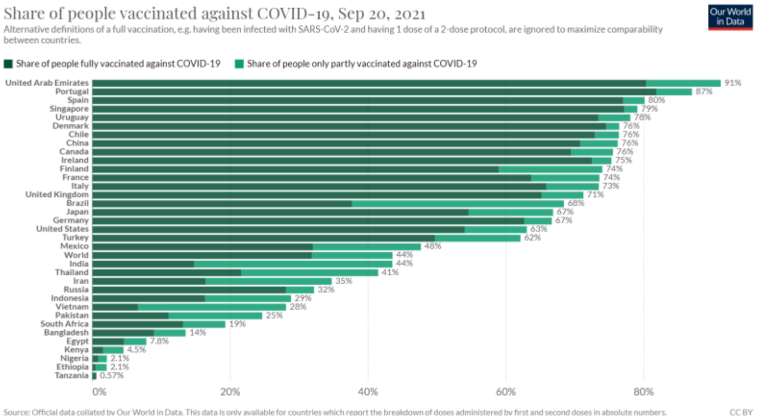
Although it might initially seem counterintuitive that older patients have lower cost of risk, this is due to the fact that the cost associated with hospitalization is heavily correlated with the time spent at a hospital and the older population has a higher concentration of deceased individuals making their stay potentially shorter.

As discussed in the appendix, a potential enhancement of our approach would be to fine-tune these figures to the realities observed in *Humana* in particular. Moreover, an analytics model could be built to estimate this expected cost of hospitalization at the individual level instead of assigning group averages from the market.

4 Determining potential interventions to enhance vaccination

In this fourth chapter we explore different potential interventions that could help enhance the vaccination intake. While the rest of our effort – quantifying the different dimensions of the intervention benefit formula – is a quantitative task, in order to have real-world actionability, the appropriate interventions need to be chosen which is somewhat more qualitative. A first glimpse at vaccination rates by country allows us to understand potential interventions in those countries where the intake has had more historical success. In exhibit 4-1 we can observe these discrepancies by country.

EXHIBIT 4-1[11]



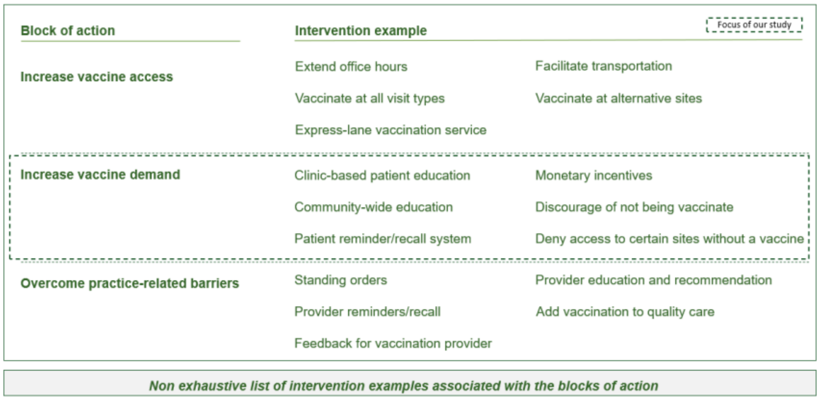
Share of people vaccinated against COVID-19, Sep 20, 2021

Alternative definitions of a full vaccination, e.g. having been infected with SARS-CoV-2 and having 1 dose of a 2-dose protocol, are ignored to maximize comparability between countries.

Source: Official data collated by Our World in Data. This data is only available for countries which report the breakdown of doses administered by first and second doses in absolute numbers. CC BY

Nonetheless, country specific circumstances might make these policies not applicable to the US, focusing on past success histories in the US, an exhaustive study named *Practice-Proven Interventions to Increase Vaccination Rates and Broaden the Immunization Season*[12] breaks down the strategies to increase coverage in three categories:

EXHIBIT 4-2



Block of action	Intervention example	Focus of our study
Increase vaccine access	Extend office hours	Facilitate transportation
	Vaccinate at all visit types	Vaccinate at alternative sites
	Express-lane vaccination service	
Increase vaccine demand	Clinic-based patient education	Monetary incentives
	Community-wide education	Discourage of not being vaccinate
	Patient reminder/recall system	Deny access to certain sites without a vaccine
Overcome practice-related barriers	Standing orders	Provider education and recommendation
	Provider reminders/recall	Add vaccination to quality care
	Feedback for vaccination provider	

Non exhaustive list of intervention examples associated with the blocks of action

[11] https://ourworldindata.org/covid-vaccinations

[12] https://www.amjmed.com/article/S0002-9343(08)00466-X/pdf

This dichotomy of the alternatives into three blocks is very illustrative – namely increasing vaccination access, increasing the demand, or overcoming practice related barriers. The focus of our study is on the second one, i.e. increasing the demand. Within the second category – increasing the demand, the CDC[13] outlines the following 6 potential cases as causes of lack of demand

- Limited access to health care
- Multiple competing priorities for providers who care for adult patients
- Low awareness among adults about recommended vaccines and their benefits
- Challenges in coordinating care for adults who often have more than one medical provider
- A complicated adult immunization schedule
- Vaccine cost and reimbursement

In the context of COVID 19, we will focus on stimulating the third one, which stands for low awareness and resistance to the vaccine or hesitancy – the scope of this study. Because of their success in some demonstrated studies[14] and their potential in others, we have focused on the *"Missed opportunity vaccination intervention guidebook from the World Health Organization"*[15] where a comprehensive list of considerations to be made before selecting which interventions to pursue is outlined.

We have defined five types of targeted interventions where in each of them the *Humana* member would receive information reminding: **(1) Urgency**: They are yet to be vaccinated – **(2) Price**: The vaccine is free of charge – **(3) Logistics**: Location and opening hours of the closest vaccination clinic to their residence. These five interventions are assigned in a two-step process, first assigning the optimal first touch point for the client (digital reminder or newsletter) and then decide if it is profitable to assign a second step intervention being – phone call, discount and cash disbursement

- **Digital reminder**: Recall based on the digital channel preference for every member (SMS, email, intranet…) conduct a targeted *Humana* member reach out
- **Newsletter education**: Send a non-intrusive newsletter through the preferred digital channel of every *Humana* member highlighting the benefits of the vaccine and dangers of not being vaccinated
- **Healthcare provider phone call**: Interact through a call highlighting the need and logistics of the vaccination
- **Discount in healthcare plan[16]**: Notify eligibility for a monetary incentive in case of vaccination. A recent study at UCLA indicated that a third of the unvaccinated population would make them more likely to get a shot
- **Cash disbursement**: For some Humana members, release a message notifying that they are eligible for a direct cash disbursement and not necessarily associated with a discount in the healthcare plan

[13] https://www.cdc.gov/vaccines/pubs/pinkbook/strat.html

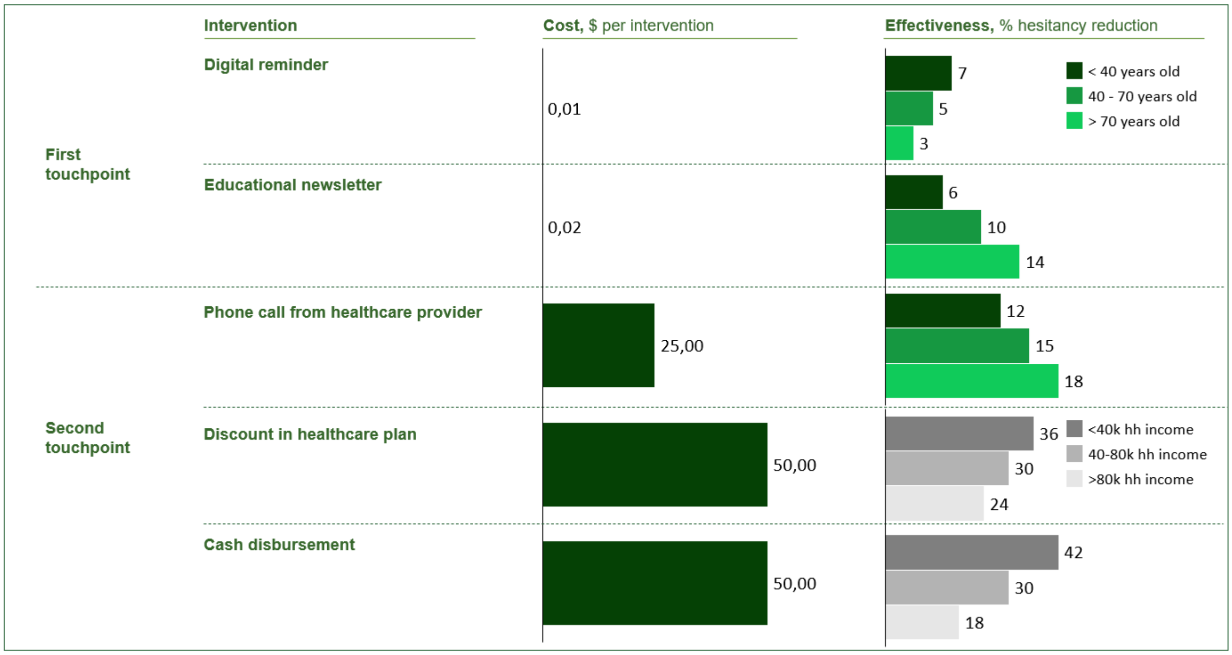
[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4514191/table/T2/?report=objectonly

[15] https://www.who.int/immunization/programmes_systems/policies_strategies/MOV_Intervention_Guidebook.pdf

[16] https://www.nytimes.com/2021/05/04/upshot/vaccine-incentive-experiment.html

In order to be able to compute the exact benefit of these interventions at a member level, we have estimated the parameters from exhibit 4-3 which characterize the interventions from a cost and effectiveness standpoint. This estimation comes from "*Strategies for addressing vaccine hesitancy*"[17] from the World Health Organization

EXHIBIT 4-3



Outside-in estimation that could be refined with more granular data from Humana's specific context

In the appendix section we discuss how this approximation could be fine-tuned to reflect the realities of *Humana* members and even fine-grained so that every member gets an individualized cost and effectiveness estimate. Although our study mainly focuses on how to best allocate resources to conduct a targeted response – that is, identifying which members will better respond to which interventions – some group-wide interventions deserve to be considered such as:

- **Run a vaccine lottery**
- **Establish partnerships with pharmacies / businesses to give immediate discounts**
- **Location speeches**
- **Commercial awareness**

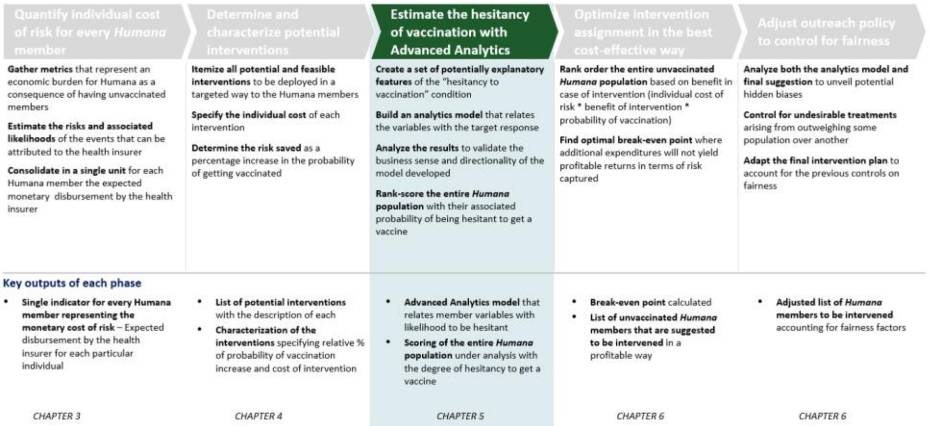
The decision to conduct a group-wide intervention – partnership with employers, conferences… - could be fueled with the indicators we are creating in this study, but it is outside the scope of the present body of work, in the appendix section we discuss about it with more detail.

[17] https://www.who.int/immunization/sage/meetings/2014/october/3_SAGE_WG_Strategies_addressing_vaccine_hesitancy_2014.pdf

5 Estimating the hesitancy of vaccination

The focus of the fifth chapter is to present an in-depth step-by-step approach to the solution followed to estimate the likelihood of not being vaccinated (defined in this context as hesitancy). It is structured as a sequence of subchapters beginning with the data gathering and consolidation process to later discuss the definition of the population and target variable. Once all the preliminary ingredients are ready, we study the algorithm developed together with the interpretation of its results and the fairness consequences. Following the framework discussed in chapter 2, this section is dedicated to study the shaded step in exhibit 5-1.

EXHIBIT 5-1



Quantify individual cost of risk for every *Humana* member

Gather metrics that represent an economic burden for Humana as a consequence of having unvaccinated members

Estimate the risks and associated likelihoods of the events that can be attributed to the health insurer

Consolidate in a single unit for each Humana member the expected monetary disbursement by the health insurer

Determine and characterize potential interventions

Itemize all potential and feasible interventions to be deployed in a targeted way to the Humana members

Specify the individual cost of each intervention

Determine the risk saved as a percentage increase in the probability of getting vaccinated

Estimate the hesitancy of vaccination with Advanced Analytics

Create a set of potentially explanatory features of the "hesitancy to vaccination" condition

Build an analytics model that relates the variables with the target response

Analyze the results to validate the business sense and directionality of the model developed

Rank-score the entire *Humana* population with their associated probability of being hesitant to get a vaccine

Optimize intervention assignment in the best cost-effective way

Rank order the entire unvaccinated *Humana* population based on benefit in case of intervention (individual cost of risk * benefit of intervention * probability of vaccination)

Find optimal break-even point where additional expenditures will not yield profitable returns in terms of risk captured

Adjust outreach policy to control for fairness

Analyze both the analytics model and final suggestion to unveil potential hidden biases

Control for undesirable treatments arising from outweighing some population over another

Adapt the final intervention plan to account for the previous controls on fairness

Key outputs of each phase

- **Single indicator for every Humana member representing the monetary cost of risk** – Expected disbursement by the health insurer for each particular individual

CHAPTER 3

- **List of potential interventions** with the description of each
- **Characterization of the interventions** specifying relative % of probability of vaccination increase and cost of intervention

CHAPTER 4

- **Advanced Analytics model** that relates member variables with likelihood to be hesitant
- **Scoring of the entire *Humana* population** under analysis with the degree of hesitancy to get a vaccine

CHAPTER 5

- **Break-even point** calculated
- **List of unvaccinated *Humana* members that are suggested to be intervened** in a profitable way

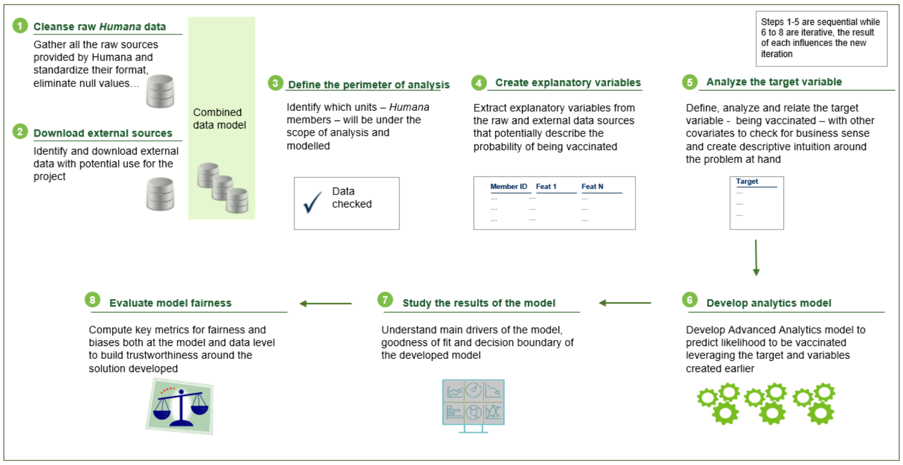
CHAPTER 6

- **Adjusted list of *Humana* members to be intervened** accounting for fairness factors

CHAPTER 6

Additionally, this fifth chapter builds some necessary intuition around the problem and analytics solution developed and extracts key insights that will be crucial to derive the implications for *Humana* discussed in chapter 6. Illustration 5-2 depicts the steps followed to bridge from the raw data to the unbiased likelihood of hesitancy estimates.

EXHIBIT 5-2



1 **Cleanse raw *Humana* data**

Gather all the raw sources provided by Humana and standardize their format, eliminate null values…

2 **Download external sources**

Identify and download external data with potential use for the project

Combined data model

3 **Define the perimeter of analysis**

Identify which units – *Humana* members – will be under the scope of analysis and modelled

✓ Data checked

4 **Create explanatory variables**

Extract explanatory variables from the raw and external data sources that potentially describe the probability of being vaccinated

Member ID	Feat 1	Feat N
…	…	…
…	…	…
…	…	…

5 **Analyze the target variable**

Define, analyze and relate the target variable - being vaccinated – with other covariates to check for business sense and create descriptive intuition around the problem at hand

Target
…
…
…

Steps 1-5 are sequential while 6 to 8 are iterative, the result of each influences the new iteration

6 **Develop analytics model**

Develop Advanced Analytics model to predict likelihood to be vaccinated leveraging the target and variables created earlier

7 **Study the results of the model**

Understand main drivers of the model, goodness of fit and decision boundary of the developed model

8 **Evaluate model fairness**

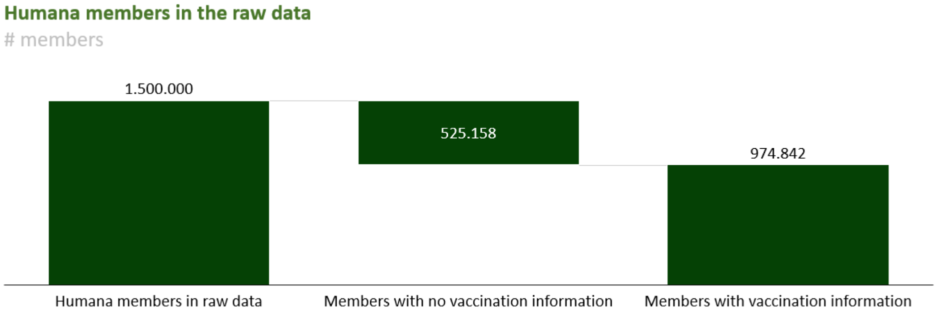
Compute key metrics for fairness and biases both at the model and data level to build trustworthiness around the solution developed

As one can note, the steps until the master table consolidation are sequential but the last refinements involve iterations since the results of the model will in part influence its posterior recalibrations to account for fairness, biases, and goodness of fit.

5.1 HUMANA'S DATA OVERVIEW

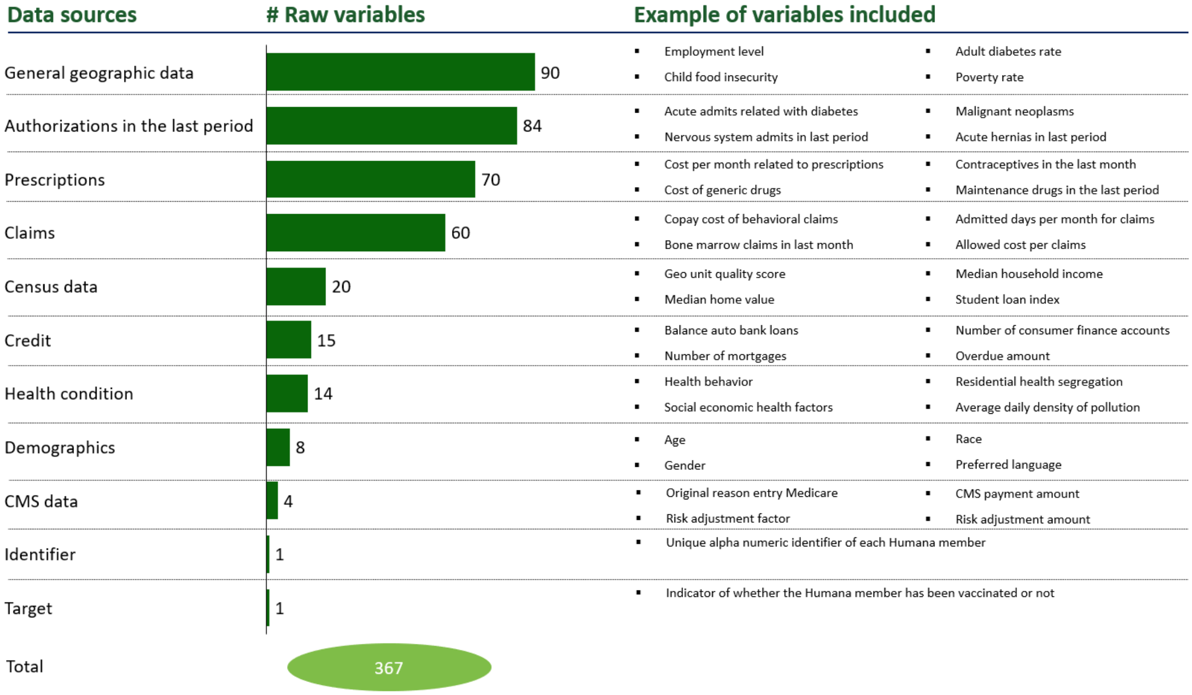
The first step in any analytics study consists of analyzing the available data, its content and format. From the raw data provided by *Humana*, we observe that there are 1.5Mn unique members, 974k of whom have the vaccination information available.

EXHIBIT 5-3



For this 1.5Mn Humana members, we have at our disposal 367 raw variables from 11 different data blocks. The breakdown of raw variables per data block is depicted in exhibit 5-4.

EXHIBIT 5-4



Data sources	# Raw variables	Example of variables included	
General geographic data	90	▪ Employment level ▪ Child food insecurity	▪ Adult diabetes rate ▪ Poverty rate
Authorizations in the last period	84	▪ Acute admits related with diabetes ▪ Nervous system admits in last period	▪ Malignant neoplasms ▪ Acute hernias in last period
Prescriptions	70	▪ Cost per month related to prescriptions ▪ Cost of generic drugs	▪ Contraceptives in the last month ▪ Maintenance drugs in the last period
Claims	60	▪ Copay cost of behavioral claims ▪ Bone marrow claims in last month	▪ Admitted days per month for claims ▪ Allowed cost per claims
Census data	20	▪ Geo unit quality score ▪ Median home value	▪ Median household income ▪ Student loan index
Credit	15	▪ Balance auto bank loans ▪ Number of mortgages	▪ Number of consumer finance accounts ▪ Overdue amount
Health condition	14	▪ Health behavior ▪ Social economic health factors	▪ Residential health segregation ▪ Average daily density of pollution
Demographics	8	▪ Age ▪ Gender	▪ Race ▪ Preferred language
CMS data	4	▪ Original reason entry Medicare ▪ Risk adjustment factor	▪ CMS payment amount ▪ Risk adjustment amount
Identifier	1	▪ Unique alpha numeric identifier of each Humana member	
Target	1	▪ Indicator of whether the Humana member has been vaccinated or not	
Total	367		

Our goal in this chapter will be to find patterns between these rich data sources and the flag of vaccination for the 974k members and then evaluate whether we have a sufficiently strong evidence to generalize for the remainder of the 1.5Mn population. As one can note, these rich sources will allow for a broader prediction, i.e. not limited to health data or patient's history but also containing other more diverse attributes.

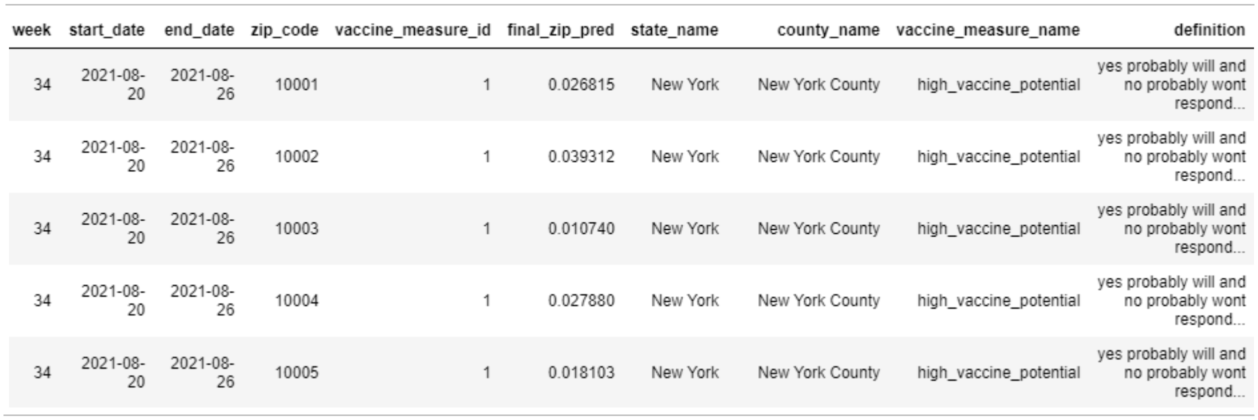
5.2 EXTERNAL DATA UTILIZED

In addition to the raw sources provided by *Humana* and detailed in the previous subsection, we have collected external and publicly available information that covers different dimensions and enriches the upcoming analyses. Namely, the additional data sources leveraged are the following:

- **Vaccination hesitancy surveys from healthdata.org[18]:** This weekly data at the zip code level offers a rich view of which geographies are more hesitant to be vaccinated. A sample in each US zip code is surveyed with the following two questions:
    - *Yes, probably will and no probably won't respondents*
    - *Yes, probably will, no probably won't, and no definitely won't respondents*

  We use this data as a structural screenshot in time to allow the model to rank-order geographies based on the degree to which people are hesitant in them. Exhibit 5-5 illustrates the raw format of this data source.

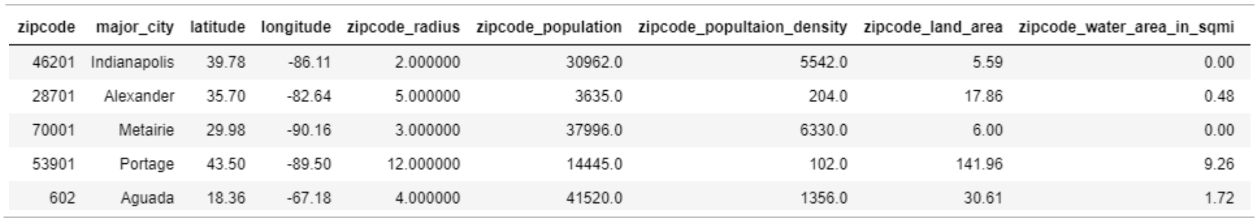
EXHIBIT 5-5



week	start_date	end_date	zip_code	vaccine_measure_id	final_zip_pred	state_name	county_name	vaccine_measure_name	definition
34	2021-08-20	2021-08-26	10001	1	0.026815	New York	New York County	high_vaccine_potential	yes probably will and no probably wont respond...
34	2021-08-20	2021-08-26	10002	1	0.039312	New York	New York County	high_vaccine_potential	yes probably will and no probably wont respond...
34	2021-08-20	2021-08-26	10003	1	0.010740	New York	New York County	high_vaccine_potential	yes probably will and no probably wont respond...
34	2021-08-20	2021-08-26	10004	1	0.027880	New York	New York County	high_vaccine_potential	yes probably will and no probably wont respond...
34	2021-08-20	2021-08-26	10005	1	0.018103	New York	New York County	high_vaccine_potential	yes probably will and no probably wont respond...

- **Zip code structural data from *uszipcode*[19]**: Additionally to the geographic data provided from *Humana*, we have leveraged a public API to retrieve zip code specific characteristics. These characteristics contain complementary data such as housing units, occupied housing units, population density, water area… which can add value in the upcoming prediction task. The raw format of this data can be found in exhibit 5-6 and in the future references on the synthetic variables utilized for the prediction task.

EXHIBIT 5-6



zipcode	major_city	latitude	longitude	zipcode_radius	zipcode_population	zipcode_popultaion_density	zipcode_land_area	zipcode_water_area_in_sqmi
46201	Indianapolis	39.78	-86.11	2.000000	30962.0	5542.0	5.59	0.00
28701	Alexander	35.70	-82.64	5.000000	3635.0	204.0	17.86	0.48
70001	Metairie	29.98	-90.16	3.000000	37996.0	6330.0	6.00	0.00
53901	Portage	43.50	-89.50	12.000000	14445.0	102.0	141.96	9.26
602	Aguada	18.36	-67.18	4.000000	41520.0	1356.0	30.61	1.72

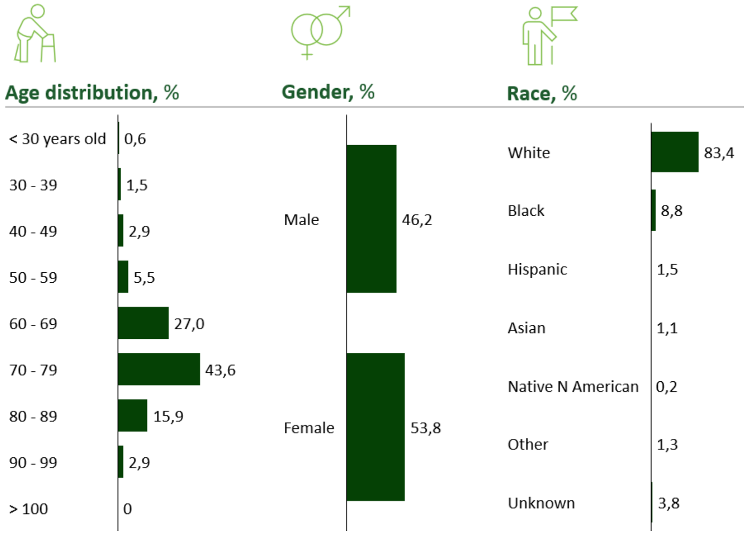
[18] https://vaccine-hesitancy.healthdata.org/ and http://www.healthdata.org/acting-data/covid-19-vaccine-hesitancy-us-county-and-zip-code

[19] https://github.com/MacHu-GWU/uszipcode-project

5.3 PERIMETER OF ANALYSIS – HUMANA MEMBERS ANALYZED

Once all the data has been gathered, cleansed, and consolidated, the immediate next step consists of defining the perimeter of analysis or population. In our study this perimeter has already been outlined by the *Humana* organizing team already, namely the 1.5Mn members who were in a Medicare advantage plan[20] and eligible to get vaccinated in March 2021. With this definition we first observe the traits that characterize this population in exhibit 5-7.

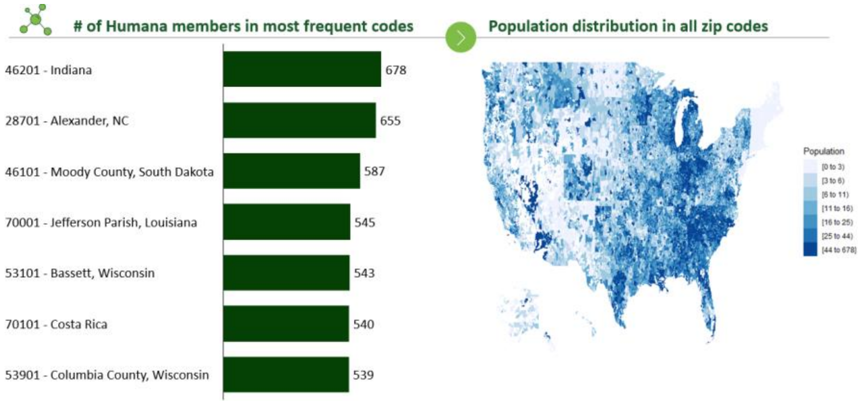
EXHIBIT 5-7



While the gender distribution appears to be balanced, both the race and age are skewed. We can see that the age distribution of the 1.5Mn Humana members is skewed to the older side with an average age of 71 years old. Similarly, 83% of all the members under analysis are white.

Analyzing the data at the zip code level, we can see how the 1.5Mn of members in our perimeter of analysis are distributed across the US. Exhibit 5-8 illustrates that there are some zip codes with high concentration of *Humana* members and taking it to the country level we see some "hot spots" that concentrate higher counts than others.

EXHIBIT 5-8



[20] https://www.medicare.gov/sign-up-change-plans/types-of-medicare-health-plans/medicare-advantage-plans

5.4 EXPLANATORY VARIABLES

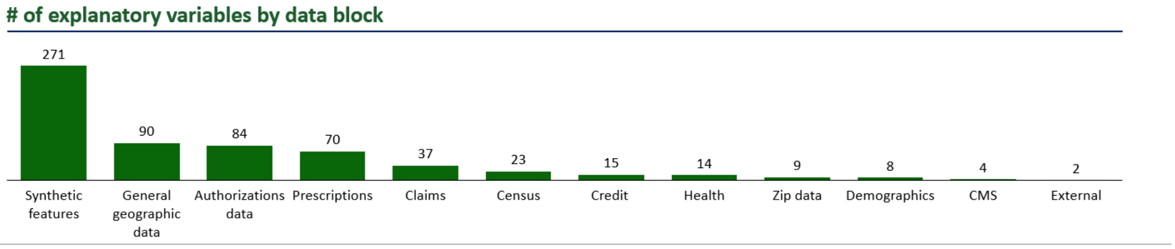
The next step in the model development phase consists of extracting explanatory variables at the member level which can be later utilized by the advanced analytics model. For each of the variables in every data source, we have proceeded doing the following:

1. **Variable cleansing**:
   - **Renaming**: Based on the definitions provided by *Humana* all the raw variables have been renamed for the ease of comprehension throughout the study
   - **Filling null values**: Most variables have different reasons of nullable values, we have carried out contextual imputation, i.e. in some cases filling by 0, others by the mean and others by the mode depending on the context of the variable at hand and the potential reason for it being null
   - **Assigning the correct variable type**: In order to have homogeneity across variables, all have been converted to the type "*float64*", this unveiled that some variables had unexpected values like "*" or text within a numeric variable and we controlled to solve these issues. The analytics model will ingest a table only filled with numeric values
   - **Label encoding of categorical values**: Similar to the previous step, all categorical variables have been mapped to numeric values, when the ordering between them may have sense it has been preserved and otherwise it is let as a label encoded feature and we specify to the algorithm which are categorical features[21]
   - **Fill missing zip code information**: For some of the external data we were not able to map to the exact zip code, we have followed a proximity approach assigning the external data of the closest zip code available

2. **Combine with other variables to create synthetic indicators**:
   - **Direct explanatory features**: These consist of the already cleansed variables with no changes
   - **Zip code aggregations**: Average of other variables (income, health risk…) by code of the member
   - **Zip code discrepancies**: Discrepancy between the variables (income) and the average of code
   - **Age quotients**: Division of some indicators against age – loans, health risk, # admissions…
   - **Count of categorical variables**: Count encoding for the categorical variables, # appearances

After these 3 steps and extracting several additional features that might be relevant for the modelling process, we have a resulting master data table that consists of 627 explanatory variables. These can be classified into 12 blocks as depicted in exhibit 5-9.

EXHIBIT 5-9



[21] https://lightgbm.readthedocs.io/en/latest/Features.html#optimal-split-for-categorical-features

Once we have cleansed and created a comprehensively exhaustive set of explanatory variables, a necessary sanity check consists of validating their disparity between training and testing data. Since our ultimate goal throughout this chapter is to find patterns between the covariates and the hesitancy response that generalizes to the test set (unseen labels), it is fundamental that the test set is similarly distributed to the train data. We have first analyzed the decile distribution of each of the 627 variables to validate that their range is consistent with their description as can be seen for some variables in exhibit 5-10.

EXHIBIT 5-10

Variable	Decile1	Decile2	Decile3	Decile4	Decile5	Decile6	Decile7	Decile8	Decile9
atlas_agritrsm_rct12	5000,00	14000,00	38000,00	79000,00	141000,00	210699,42	210699,42	210699,42	463000,00
atlas_avghhsize	2,37	2,42	2,47	2,51	2,55	2,59	2,64	2,72	2,78
atlas_berry_acrespth12	0,00	0,01	0,04	0,08	0,16	0,21	0,21	0,21	0,36
atlas_berry_farms12	0,00	2,00	3,00	4,00	6,00	8,00	10,00	14,00	25,00
atlas_convspth14	0,27	0,33	0,36	0,40	0,45	0,49	0,55	0,62	0,72
atlas_csa12	0,00	1,00	1,00	2,00	3,00	3,00	4,00	6,00	8,00
atlas_deep_pov_all	4,09	4,99	5,68	6,31	6,87	7,33	7,69	8,19	9,19
atlas_deep_pov_children	4,64	6,00	7,10	8,04	9,09	9,90	10,72	11,86	14,47
atlas_dirsales_farms12	12,00	18,00	28,00	35,00	42,00	49,00	61,00	84,00	120,00
atlas_farm_to_school13	0,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
atlas_ffrpth14	0,43	0,52	0,57	0,63	0,67	0,71	0,75	0,79	0,84
atlas_fmrktpth16	0,00	0,01	0,01	0,02	0,02	0,03	0,03	0,04	0,06
atlas_foodhub16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_foodinsec_13_15	11,10	12,20	12,70	13,80	14,60	14,90	15,40	16,10	17,60
atlas_foodinsec_child_03_11	7,40	7,80	8,00	8,70	9,23	9,40	10,00	10,50	11,40
atlas_freshveg_farms12	3,00	6,00	10,00	13,00	15,00	17,00	21,00	29,00	47,00
atlas_fsrpth14	0,42	0,52	0,58	0,63	0,69	0,73	0,77	0,84	0,96
atlas_ghveg_farms12	0,00	0,00	1,00	1,00	2,00	3,00	5,00	6,00	9,00
atlas_ghveg_sqftpth12	0,00	0,00	0,00	0,22	15,68	60,48	60,48	60,48	137,80
atlas_grocpth14	0,11	0,13	0,14	0,15	0,16	0,18	0,19	0,21	0,25
atlas_hh65plusalonepct	8,36	9,25	10,11	10,51	11,00	11,48	12,06	12,92	14,11
atlas_hiamenity	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	1,00
atlas_hipov_1115	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00
atlas_low_education_2015_update	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_low_employment_2015_update	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00
atlas_medhhinc	39704,00	42998,00	46239,00	49277,00	51549,00	54053,00	56855,00	60025,00	67503,00
atlas_naturalchangerate1016	-1,02	0,01	0,60	1,19	1,73	2,31	3,04	3,66	4,46
atlas_net_international_migration_rate	0,07	0,16	0,25	0,36	0,48	0,66	1,01	1,44	2,11
atlas_netmigrationrate1016	-3,33	-1,84	-0,96	-0,01	0,91	1,82	3,19	4,57	6,57
atlas_orchard_acrespth12	0,01	0,09	0,17	0,31	0,54	0,80	1,25	2,03	2,79
atlas_orchard_farms12	2,00	4,00	6,00	9,00	11,00	15,00	19,00	26,00	46,00
atlas_ownhomepct	56,67	62,60	65,74	68,67	70,57	72,16	73,77	75,78	78,40
atlas_pc_dirsales12	0,16	0,45	0,80	1,38	2,30	3,30	4,24	6,00	8,97
atlas_pc_ffrsales12	495,41	523,89	567,30	608,73	622,68	642,49	649,09	665,32	677,73
atlas_pc_fsrsales12	557,09	598,47	623,59	642,78	649,84	688,25	690,52	738,37	882,94
atlas_pc_snapben15	8,94	11,63	13,93	15,75	17,46	18,64	20,34	23,27	27,42
atlas_pc_wic_redemp12	11,20	13,15	15,00	16,81	18,44	18,52	20,44	22,57	26,71
atlas_pct_cacfp15	0,82	0,99	1,10	1,13	1,24	1,33	1,39	1,64	1,73
atlas_pct_diabetes_adults13	8,50	9,20	9,80	10,30	10,80	11,40	12,10	12,70	13,80
atlas_pct_fmrkt_anmlprod16	0,00	33,33	50,00	50,00	55,98	62,63	70,00	87,50	100,00
atlas_pct_fmrkt_baked16	0,00	45,56	50,00	58,18	60,32	66,67	75,00	100,00	100,00
atlas_pct_fmrkt_credit16	0,00	0,00	33,33	49,88	50,00	60,00	70,00	80,00	100,00
atlas_pct_fmrkt_frveg16	15,00	50,00	52,00	62,58	64,67	72,23	81,15	100,00	100,00
atlas_pct_fmrkt_otherfood16	0,00	33,33	50,00	55,01	57,72	66,67	75,00	100,00	100,00
atlas_pct_fmrkt_sfmnp16	0,00	0,00	0,00	0,00	0,00	4,04	17,32	35,23	60,67
atlas_pct_fmrkt_snap16	0,00	0,00	0,00	0,00	8,33	21,19	28,17	41,60	60,00
atlas_pct_fmrkt_wic16	0,00	0,00	0,00	0,00	0,00	0,00	13,67	28,88	50,00
atlas_pct_fmrkt_wiccash16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	4,56	17,29

Next, and in order to check for structural differences between train and test sets, we have taken the differences of the deciles of each variable between train and test. High values in these would yield a difference not only at an average level but at a distribution level, this has served us to further cleanse the variables, exhibit 5-11 depicts an example of how the distribution of the difference looks like for some of the variables.

EXHIBIT 5-11

Var	Decile1	Decile2	Decile3	Decile4	Decile5	Decile6	Decile7	Decile8	Decile9
atlas_avghhsize	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_berry_acrespth12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_berry_farms12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-1,00	0,00
atlas_convspth14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_csa12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_deep_pov_all	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_deep_pov_children	0,00	0,00	0,00	-0,01	0,00	0,00	-0,01	0,00	0,00
atlas_dirsales_farms12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_farm_to_school13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_ffrpth14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_fmrktpth16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_foodhub16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_foodinsec_13_15	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_foodinsec_child_03_11	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_freshveg_farms12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_fsrpth14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_ghveg_farms12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
atlas_ghveg_sqftpth12	0,00	0,00	0,00	0,05	0,00	0,00	0,00	0,00	2,78
atlas_grocpth14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

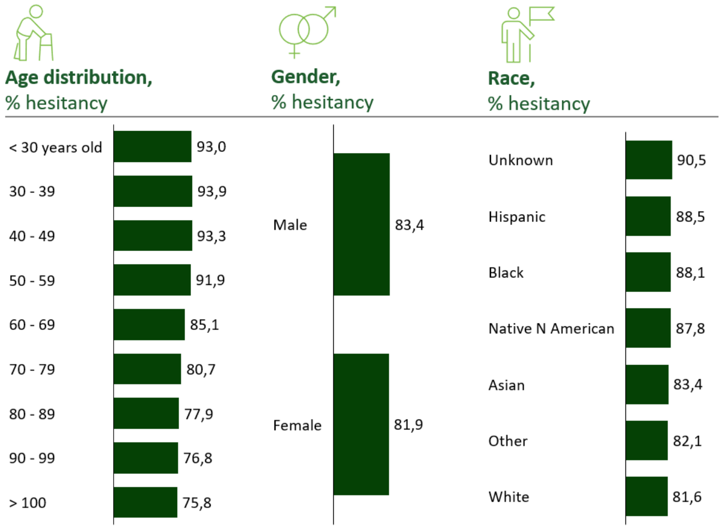
5.5 TARGET VARIABLE DEFINITION – HESITANCY OF VACCINATION INDICATOR

The target response is the variable that we want to learn patterns from and extract predictions for the hold out test. As discussed earlier, our focus will be in estimating the likelihood of being hesitant to get a vaccine and with this purpose the data facilitated from Humana already has this flag at the member level. For every member in the *Humana* population provided we know if they got the vaccine or not.

While this prediction should be performed in a rolling window basis, i.e. utilize the last batch of N months available data to see if the individual gets a vaccine in the next M months, in the context of this competition we only have at our disposal static data from a particular point in time.

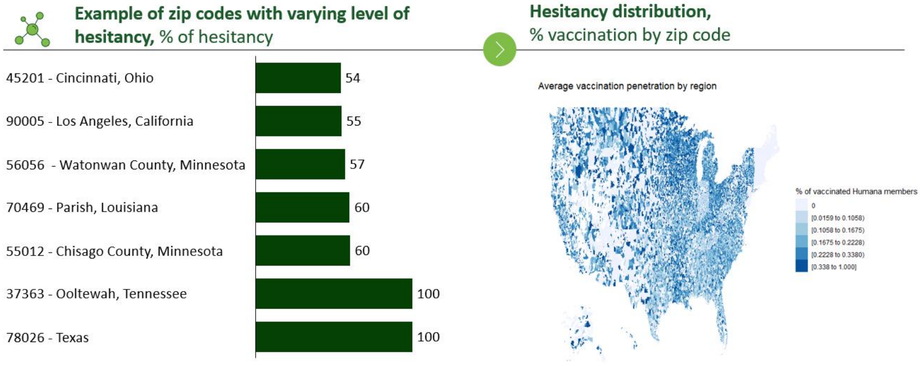
Once the target for the advanced analytics model has been defined, namely the flag of hesitancy, we can cross relate it with other covariates to test its business sense. The following exhibits try to assess and illustrate its adequacy. From exhibit 5-12 we can observe that Humana members with certain demographic characteristics show different rates of hesitancy to vaccinate. For instance, the older the member the less hesitant it is to vaccinate.

EXHIBIT 5-12



Additionally, the rate of hesitancy to vaccinate is not homogeneous across regions of the US. We can observe the contrast of hesitancy to vaccinate by state and observe that there are some with very varying levels of vaccination penetration.

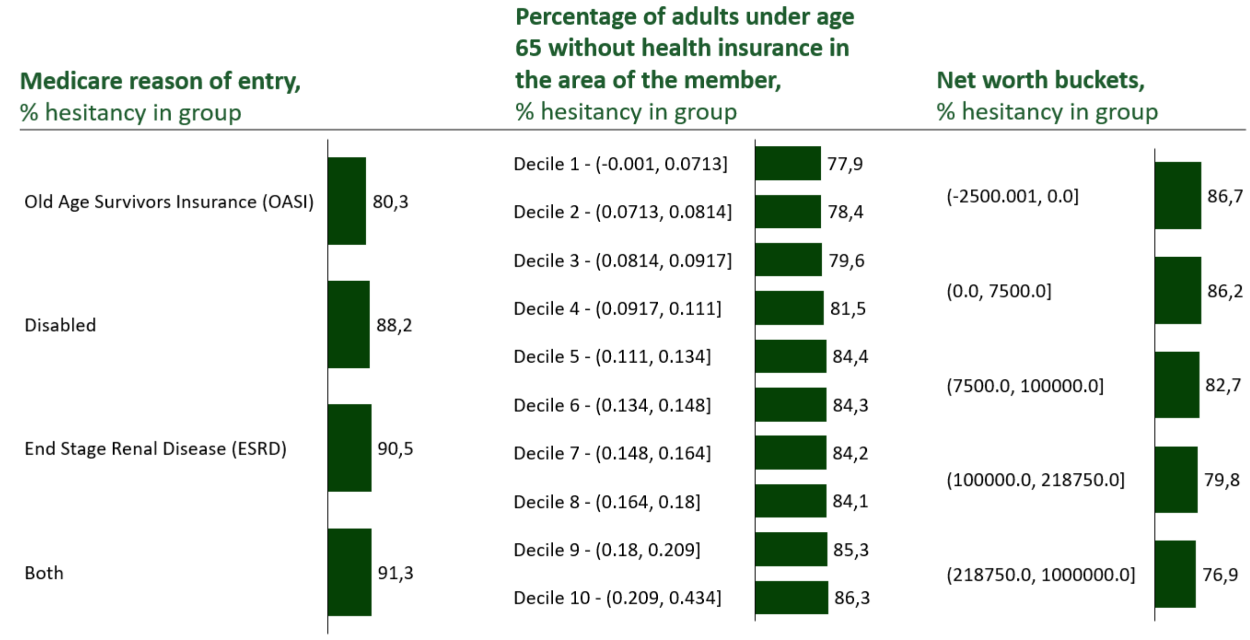
EXHIBIT 5-13


Example of zip codes with varying level of hesitancy, % of hesitancy
45201 - Cincinnati, Ohio 54
90005 - Los Angeles, California 55
56056 - Watonwan County, Minnesota 57
70469 - Parish, Louisiana 60
55012 - Chisago County, Minnesota 60
37363 - Ooltewah, Tennessee 100
78026 - Texas 100
Hesitancy distribution,
% vaccination by zip code
Average vaccination penetration by region
% of vaccinated Humana members
0
[0.0159 to 0.1058)
[0.1058 to 0.1675)
[0.1675 to 0.2228)
[0.2228 to 0.3380)
[0.338 to 1.000]

Interestingly, there are some variables that descriptively display a relation with the hesitancy of vaccination. At this point we cannot yet assess neither the predictive power nor the causality between them, but checking the individual relations builds a better understanding of the problem at hand.

For instance, we can observe that the reason of entry to the Medicare system displays some disparities across groups. Similarly, the higher the percentage of adults without health insurance in the area where the member resides yields higher levels of hesitancy. Interestingly, the higher the net worth of the Humana member, the lower the hesitancy to get a vaccine.

EXHIBIT 5-14

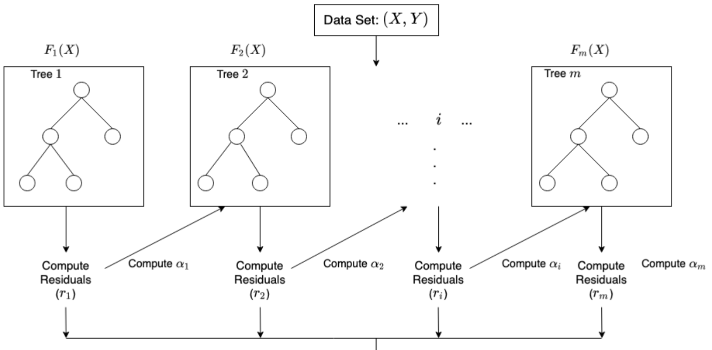


While we could do this exercise for all the variables that show a slight resemblance with the response target, we will not analyze the individual descriptive relations and contributions until after the model development in the immediate next section. The reason behind this choice is that the model will autonomously point us towards those indicators that display such a noteworthy relation with the hesitancy response.

5.6 ADVANCED ANALYTICS MODEL DEVELOPMENT

Once the population has been defined, the explanatory variables extracted and the target variable analyzed, we are ready to begin with the model development phase. Because of its flexibility, predictive power, ease of use and explainability we have mainly focused on tree-based boosting algorithms to estimate the probability of vaccination. More precisely, after trying different extreme gradient boosting frameworks, we have selected LightGBM's[22] native implementation because of the results obtained. Although we will not expand on how this family of algorithms work, exhibit 5-15 portrays the basic idea behind its steps.

EXHIBIT 5-15



$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}),$$

where α_i, and r_i are the regularization parameters and residuals computed with the i^{th} tree respectfully, and h_i is a function that is trained to predict residuals, r_i using X for the i^{th} tree. To compute α_i we use the residuals computed, r_i and compute the following: $arg \min_{\alpha} = \sum_{i=1}^{m} L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1}))$ where $L(Y, F(X))$ is a differentiable loss function.

There are three considerations that need to be made before beginning the modeling phase:

- **Metric of interest**: In this case the metric of interest is area under the receiver operating characteristic[23] as specified by the organization of the case competition. Defining this metric allows us to discriminate between models based on which one performs best in a given metric

- **Parameter fine-tuning**: This step consists of finding a set of parameters that in some sense maximizes the previously defined metric. To choose an adequate set of parameters we have executed a Bayesian Optimization algorithm[24] to find a locally maximizing set in a 5-fold cross validation framework

- **Train-validation-test split**: Once the metric of interest and parameters have been defined, the train/validation/test split needs to be established. This has more relevance in time-dependent problems than in our static setup. We have proceeded with a 5-fold cross validation framework to produce out of sample predictions for the entirety of the population at hand and the test sets for the public leaderboard

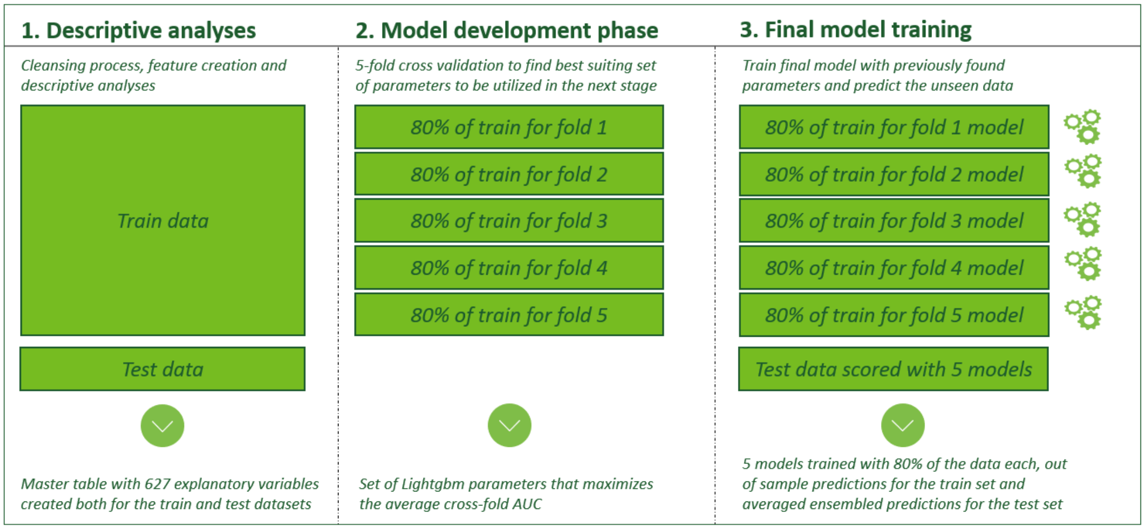
[22] https://github.com/microsoft/LightGBM

[23] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[24] https://github.com/fmfn/BayesianOptimization

Before delving into the result analysis, we will illustrate two of the three design dimensions previously explained, we won’t deep dive on the metric of interest since it is already prescribed by the competition organizers and there is no room to maneuver. The train / validation / test schema followed can be observed in the following exhibit.

EXHIBIT 5-16



1. Descriptive analyses

Cleansing process, feature creation and descriptive analyses

Train data

Test data

Master table with 627 explanatory variables created both for the train and test datasets

2. Model development phase

5-fold cross validation to find best suiting set of parameters to be utilized in the next stage

80% of train for fold 1

80% of train for fold 2

80% of train for fold 3

80% of train for fold 4

80% of train for fold 5

Set of Lightgbm parameters that maximizes the average cross-fold AUC

3. Final model training

Train final model with previously found parameters and predict the unseen data

80% of train for fold 1 model

80% of train for fold 2 model

80% of train for fold 3 model

80% of train for fold 4 model

80% of train for fold 5 model

Test data scored with 5 models

5 models trained with 80% of the data each, out of sample predictions for the train set and averaged ensembled predictions for the test set

And regarding the choice of parameters, as described earlier, we have executed a Bayesian Optimization search of parameters to maximize the 5-fold cross validation average of the AUC. That is, we have split the training the data in 5 equally sized groups, initiated 20 random sets of parameters and then performed a Bayesian search of 5 steps from the previous results (based on the relation from the parameters and the AUC found earlier the Bayesian steps to new parameter space are different). With this procedure we obtain 25 combinations of parameters – AUC with one of them maximizing it as can be seen below. The first row being the finally chosen parameters.

EXHIBIT 5-17

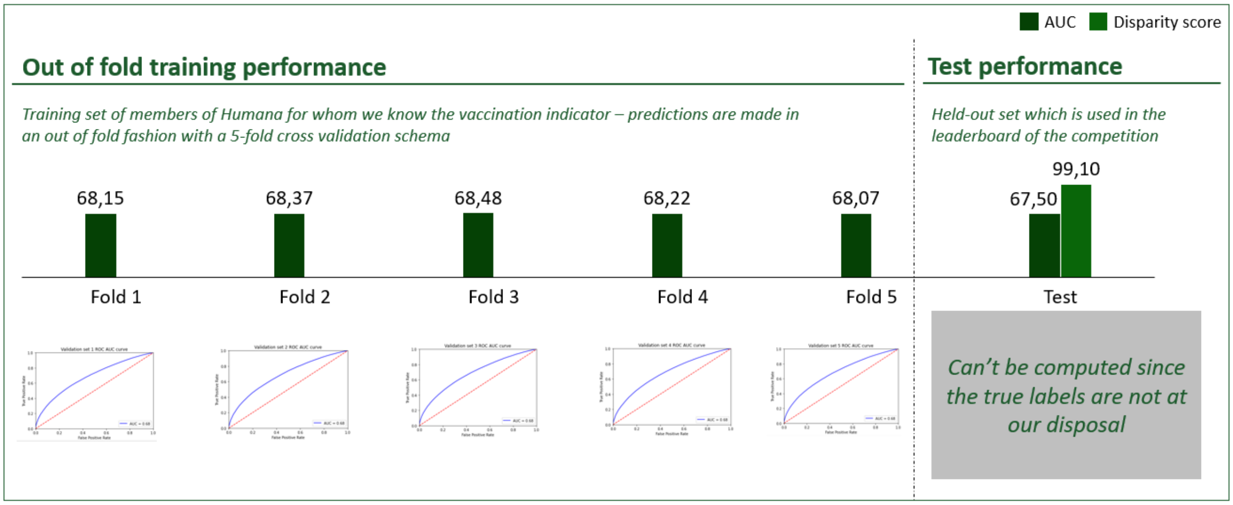
Fold AUC avg	Bagging fraction	Feature fraction	Lambda l1	Lambda l2	Maximum depth	Minimum child weight	Min split gain	Number of leaves
0,6822	0,9286	0,2211	2,5744	2,7055	6,9236	40,9941	0,0265	39,7328
0,6809	0,9094	0,3069	0,8732	1,0821	6,8224	22,5085	0,0477	44,0648
0,6807	0,9412	0,2973	1,2800	0,0720	6,2834	18,5196	0,0644	24,6662
0,6800	0,8773	0,1678	2,8100	1,9442	13,6051	13,5159	0,0955	16,7707
0,6784	0,8291	0,5114	2,6383	0,9157	7,0758	31,8610	0,0113	32,0390
0,6764	0,8801	0,5361	3,2096	1,4209	12,3265	31,5178	0,0320	37,7406
0,6762	0,8144	0,5067	2,6029	2,8051	14,5742	44,5679	0,0073	39,7840
0,6762	0,9714	0,4812	4,3157	1,9827	12,1565	40,9907	0,0322	16,1626
0,6760	0,9956	0,5816	4,1765	0,2561	11,3478	45,5141	0,0918	21,5065
0,6757	0,9177	0,5332	1,2138	0,8804	13,8319	20,8340	0,0333	33,9449
0,6756	0,9423	0,7037	0,3986	0,0462	7,1902	6,7356	0,0701	31,9446
0,6753	0,9276	0,4946	2,4589	2,3885	15,4257	9,3921	0,0296	20,0991
0,6752	0,9055	0,6810	4,3230	2,8077	8,6657	42,3061	0,0786	19,2576
0,6750	0,8007	0,5893	1,7269	1,4364	17,7778	13,7748	0,0331	44,3028
0,6747	0,9714	0,6568	2,7651	2,8057	11,6641	12,9925	0,0541	23,8038
0,6745	0,8550	0,5853	2,0581	1,1063	17,4146	5,2342	0,0956	15,9803
0,6744	0,9760	0,7270	1,2624	2,7813	10,7818	21,9718	0,0895	37,6216
0,6743	0,9757	0,6381	0,0344	0,1133	14,4615	34,1437	0,0545	23,9371
0,6737	0,9617	0,7368	1,3996	1,3663	10,1694	47,9044	0,0356	16,7828
0,6733	0,9549	0,8032	2,4074	0,9087	10,7581	28,7660	0,0612	30,7573
0,6731	0,8021	0,8070	3,2821	2,8268	14,6843	17,0245	0,0368	30,7921
0,6730	0,8671	0,8418	1,4522	0,1415	9,4920	25,8249	0,0518	20,5584
0,6728	0,8371	0,8338	1,3546	0,8206	17,4074	10,7202	0,0750	15,1571
0,6726	0,8227	0,8796	3,6437	1,0544	14,1989	40,9822	0,0649	27,4380
0,6722	0,9758	0,8854	0,1235	1,2147	16,3063	19,7048	0,0861	25,2199

5.7 RESULT ANALYSIS

In this subsection we will delve into the results obtained from the modelling effort. These consist of 4 key indicators that allow us to understand the quality and nature of our model – goodness of fit, score calibration, feature contributions, and uplift. While only the final results are displayed in this document, these metrics serve as the basis to select from one model to another in the development phase that was conducted before reaching the final model here presented.

1. **Goodness of fit**: The first result analysis to carry out is the goodness of fit check – i.e. how faithful are the model predictions with the underlying true labels. In the following exhibit one can observe the two main metrics of interest – AUC and disparity score – for the different sets of data that we have access to. While the training performance allows us to estimate whether patterns are found successfully, the other serve as guardrails to check if indeed these patterns are held in unseen data.

EXHIBIT 5-18



Out of fold training performance

Training set of members of Humana for whom we know the vaccination indicator – predictions are made in an out of fold fashion with a 5-fold cross validation schema

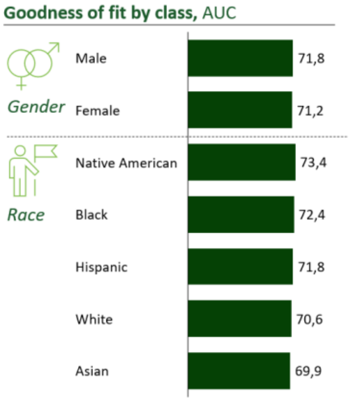
Test performance

Held-out set which is used in the leaderboard of the competition

Can't be computed since the true labels are not at our disposal

Additionally, we can double click on certain populations of interest to see how well our predictions are to group-specific individuals. In exhibit 5-19 we can observe the goodness of fit for the two variables of interest, gender, and race.

EXHIBIT 5-19



2. **Score calibration**: While in some classification use-cases a rank-ordering of the members might suffice, since we are utilizing the model scores as true odds of hesitancy to later be used as probabilities in the benefit formula, they should be aligned with the true odds coming from the data. That is, if we binarize the scoring distribution from the train data, we should see similar target averages. We analyze this behavior in the following exhibit to conclude that the scores as-are are well enough calibrated and no further modifications like isotropic regression need to be made.

EXHIBIT 5-20

Decile of model probability	Avg model probability	% Hesitant population
Decile 1 - (0.164-0.702]	64,77%	65,45%
Decile 2 - (0.702-0.748]	72,65%	72,67%
Decile 3 - (0.748- 0.781]	76,48%	76,24%
Decile 4 - (0.781-0.809]	79,50%	79,13%
Decile 5 - (0.809-0.835]	82,20%	81,75%
Decile 6 - (0.835-0.861]	84,78%	84,35%
Decile 7 - (0.861-0.886]	87,34%	87,13%
Decile 8 -(0.886-0.913]	89,95%	89,98%
Decile 9 - (0.913-0.941]	92,66%	92,80%
Decile 10 - (0.941- 0.998]	96,06%	96,68%

3. **Feature contributions**: There are several ways we can use to assess the strength with which a boosting model relies on a particular feature. During the model development (to engineer new features) and in the final model understanding, we have focused on 4 – cover, gain, sum of SHAP[25] values and sum of absolute SHAP values. Separately these metrics report on different dimensions with which features are being used. We observe that some of the top contributing factors are Age, region of the *Humana* member and risk adjustment from cms.

EXHIBIT 5-21

Feature	Cover	Gain	Sum SHAP	Sum Abs SHAP
Age of the *Humana* member	677	236.981	3.616	79.831
Categorical variable of the region	1.916	119.805	85	52.191
Age contrast of the member vs the zip code age average	637	125.383	679	45.880
atlas_pct_cacfp15	853	51.817	- 715	38.172
cms_medicare_riskadjusta	325	41.614	- 566	35.670
Categorical variable of the zip code	5.008	389.394	1.177	35.619
syn_uninsuredchild_to_age	387	48.757	1.011	31.138
prescrip_rx_gpi2_39_pmpm_cost_t_6-3-0m_b4_categorical	379	23.112	1.559	30.354
census_geounit_qualityscore	384	29.393	257	28.324
census_rx_adherence_maint	209	40.420	- 1.330	28.122
demo_mapd_behavioralsegment_categorical	903	23.995	291	27.891
prescrip_rx_bh_pmpm_ct_0to3m_b4	331	13.909	1.980	27.395
cms_totalpartypayment	708	23.307	- 472	26.736
syn_entryreason_zipcode_contrast	161	68.364	- 1.510	25.881
syn_riskadjusta_to_age	323	25.961	- 948	24.978
census_household_investableassets	129	43.051	1.185	24.004
atlas_vlfoodsec_13_15	452	44.807	- 327	23.901
syn_uninsuredadult_to_age	340	66.646	499	23.612
syn_uninsuredadult_zipcode	322	53.316	- 389	23.452
credit_hh_nonmtgcredit_60dpd	430	31.373	253	22.550

Interestingly we see some of the synthetically created features appearing as contributing such as the contrast of the age of the member against the average age in their zip code area.

[25] https://github.com/slundberg/shap

And analyzing the group-wise variable contribution, exhibit 5-22, we observe that the main feature contribution in terms of gain comes from the synthetic features derived from all the groups followed by demographic and prescriptions data:

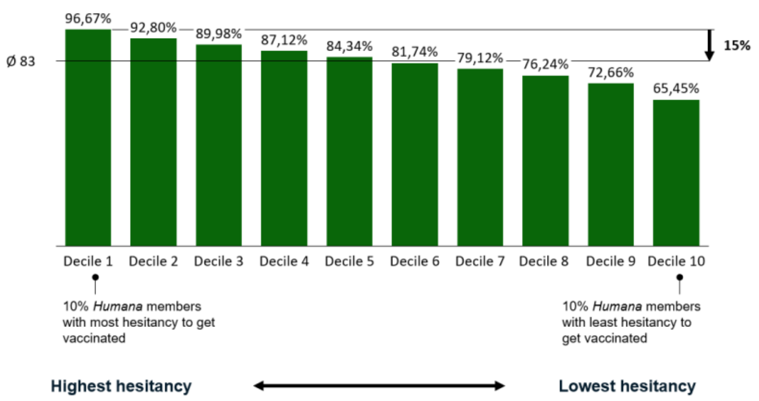
EXHIBIT 5-22

Feature group	Relative gain contribution
Synthetic features	34,16%
Demographic data	19,45%
Prescriptions	16,01%
Geographical data (atlas)	14,71%
Census data	5,75%
Credit data	3,51%
CMS data	2,19%
Health data	1,98%
Medical claims	1,69%
Zip data	0,36%
External sources	0,10%
Authorizations	0,08%

4. **Uplift**: Once the relations between the covariates and the score response from the model have been analyzed, the immediate next step is to benchmark the model performance against a random classification. While the AUC is a good proxy for this, uplift has more business meaning attached to it. In exhibit 5-24 we observe the uplift, or in other words, the factor gain that we accomplish by utilizing the model instead of a random assignment. If we were to use random assignment, we would expect to have 83% of hesitants in the first decile of likelihood to be hesitant but utilizing the model, in the first decile by model score we get 96,6% of hesitants – a 15% uplift.

EXHIBIT 5-24

Uplift curve, % of hesitancy by decile of model score



By only targeting the top 10% of *Humana* members, we would observe a 15% increase in hesitant capture against a random classification

5.8 FAIRNESS DIAGNOSTIC

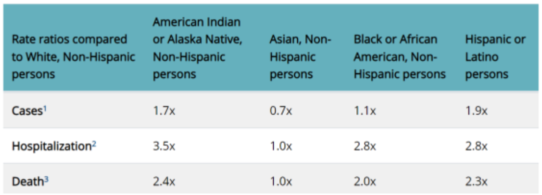
As highlighted in the case competition's guideline, an inherent concern of both Advanced Analytics models and policy making decision lies in the fairness discussion. A common definition[26] is: "*Algorithmic bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others*". Fairness presents itself in many different and often challenging formats making it difficult for practitioners and decision makers to mitigate it.

In the context of targeted vaccination outreach, examples where potential sources of unfairness can arise are:

- **Historical vaccination availability**: If only certain groups have had historical access to the vaccine, the model will rarely assign high probabilities of taking the vaccine to a minority group member
- **Disparities in the population of *Humana***: If vulnerable groups are underrepresented, the model could behave ill-manneredly for them
- **Survivorship bias**: If the data only contains members who survived through a particular period of time certain minorities that are at-risk and deceased with higher proportions could be omitted

The CDC published in September of 2021 a contrast of COVID effects broken down by ethnicity, one of the minority segmentations of interest, as can be seen in exhibit 5-26. Although the exhibit could suggest that it is worthy to invest only in higher risk populations, these figures can be biased by a large number of reasons ranging from historical access to the vaccine to propensity to self-report certain conditions among different ethnicities.

EXHIBIT 5-26[27]



Rate ratios compared to White, Non-Hispanic persons	American Indian or Alaska Native, Non-Hispanic persons	Asian, Non-Hispanic persons	Black or African American, Non-Hispanic persons	Hispanic or Latino persons
Cases[1]	1.7x	0.7x	1.1x	1.9x
Hospitalization[2]	3.5x	1.0x	2.8x	2.8x
Death[3]	2.4x	1.0x	2.0x	2.3x

To illustrate the difficulty and richness of the problem we highlight below three classic and striking fairness scenarios:

- **Simpson's paradox**: A study conducted at Berkeley[28] showed that aggregate data on admissions was bias against women, but when the data was disaggregated to the department level the bias was reversed
- **Shared confounder**: By law, loan application engines can't discriminate based on certain criteria set by GDPR or equivalent organizations – gender, race, ethnicity, zip code… However, these can often be learned by the model through common confounders[29]
- **Unequal sample rate**: Statistical parity between model results for the protected classes are often the goal of fairness analysis but sometimes are purely nonsensical – think of incidence of breast cancer by gender[30]

[26] https://en.wikipedia.org/wiki/Algorithmic_bias

[27] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html

[28] Peter J Bickel, Eugene A Hammel, J William O'Connell, and others. 1975. Sex bias in graduate admissions: Data from Berkeley. Science 187, 4175 (1975), 398–404

[29] https://auai.org/uai2019/proceedings/papers/213.pdf

[30] https://arxiv.org/pdf/1809.09245.pdf

While a unified framework to deal with bias tradeoffs has not yet been established, most state-of-the-art fairness literature focuses on studying the definition of key metrics[31]. These metrics measure disparities between protected classes and serve as proxies to inform about the trustworthiness and soundness of studies.

However, reducing the problem from a global fairness mitigation to a metric calibration is not entirely precise as most studies are significantly metric-sensitive. Some of the most commonly used techniques to reduce the metric disparities among groups are Unawareness, Demographic Parity, Equalized Odds, Predictive Parity Rate, Individual Fairness and Counterfactual Fairness[32].

We will cover two main bias sources that relate with the vaccination outreach problem – intrinsic biases from data and algorithm selection respectively. This taxonomy is consistent with the structure that the literature infers on data biases.

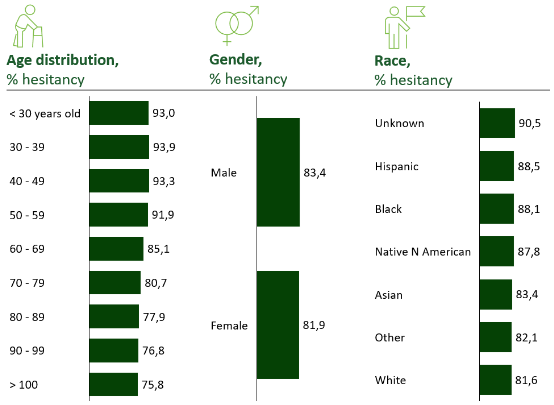
5.8.1 Data biases

To mitigate potential discriminatory outputs, the first block to analyze are the disparities coming from the raw data. These can be classified into two main groups:

- **Sample bias**: This type of bias refers to the difference in counts of individuals in the data at hand broken down by different minority classes. As we saw in the exploratory data analysis section, while the split of male / females seems to be balanced, for the race split it is predominantly whites. To conclude whether this bias is discriminatory we would have to go back to the data generation process to see if some members were being systematically excluded and analyze the outcomes – next subchapter.

- **Label bias**: This bias refers to discrepancies in the outcome variable modelled across groups of interest. Again, as noted earlier we saw that some groups display distinctively higher levels of hesitancy than others as seen in exhibit 5-27.

EXHIBIT 5-27



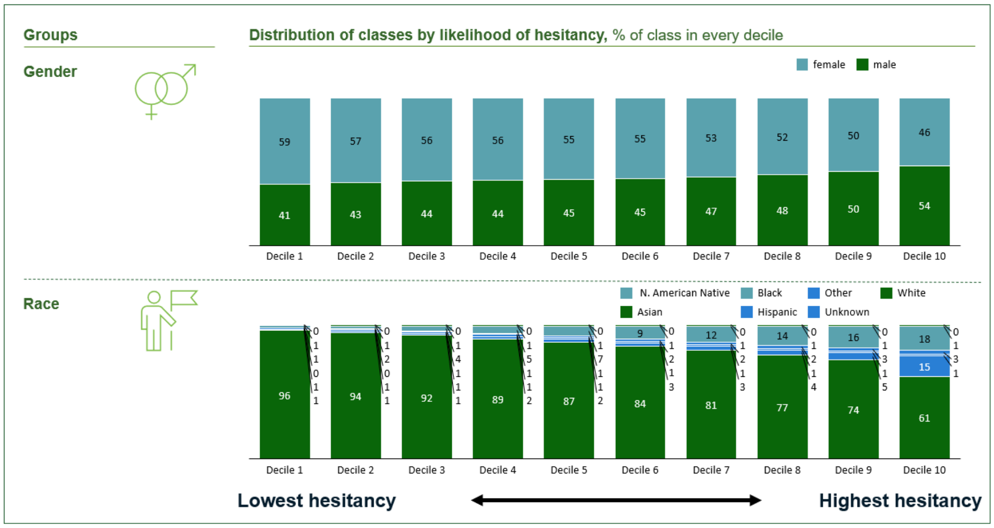
[31] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, 05 (2014), 582– 638

[32] https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

5.8.2 Model biases

In addition to the data biases explored earlier, it is worth mentioning the biases that arise from the model development as well. To determine whether sample bias or label bias are indeed influencing the outcome of our analytics effort we can refer to the distribution of classes by model likelihood. We do this with two key analyses, analyzing key metrics for every group of interest – as seen in exhibit 5-19 and by analyzing the percentage of groups in every decile of model likelihood as in exhibit 5-28.

EXHIBIT 5-28



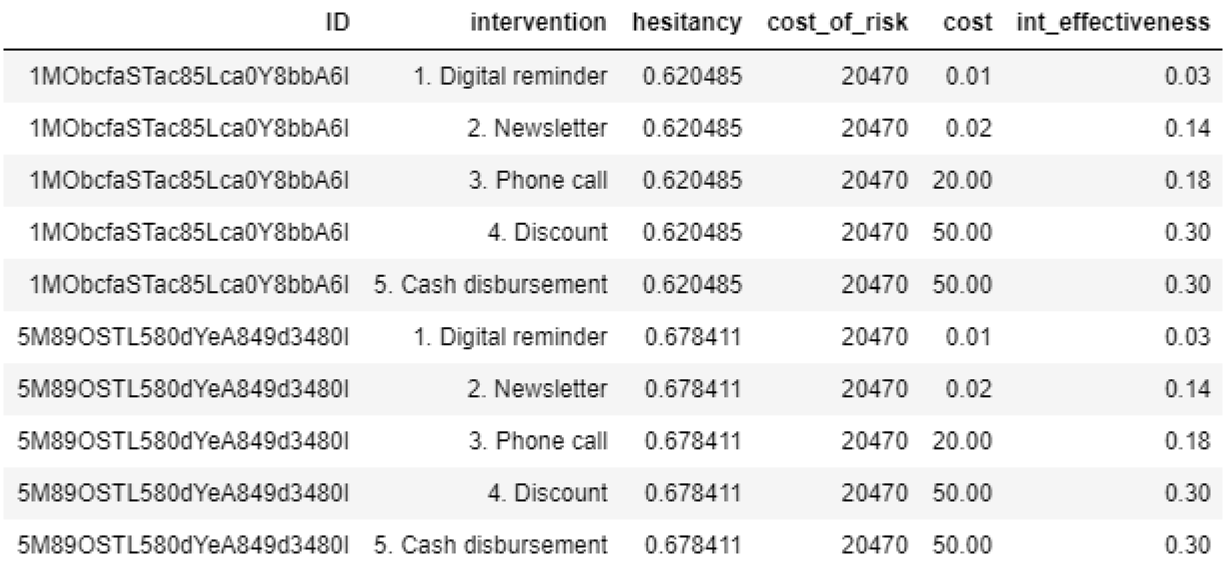
We note that the model assigns high likelihood of hesitancy to more white males than any other race or gender but in contrast with the average appearance of these classes this should come as no surprise. Interestingly, we see the percentage of whites decreasing as we increase the likelihood of hesitancy, indicating that there are other races with higher hesitancy and the model is capturing it accordingly.

Since the discrepancies do not seem significant across groups and all minority groups are represented in the first likelihood decile by the model, we will not delve deeper into how to tweak the model / data in this study but in the appendix, we highlight some techniques that could help accomplish more balance.

6 Optimizing the intervention assignment

Having at hand the key indicators calculated in the previous chapters, in this sixth section we bridge from the theoretical discussion to the applied field. More specifically, we outline how *Humana* can directly capture value from these indicators by creating an optimal targeted outreach plan. Exhibit 6-1 illustrates the resulting dataframe from the previous efforts. For every pair of member-intervention we have the cost of risk, hesitancy, cost of interventions and intervention effectiveness.

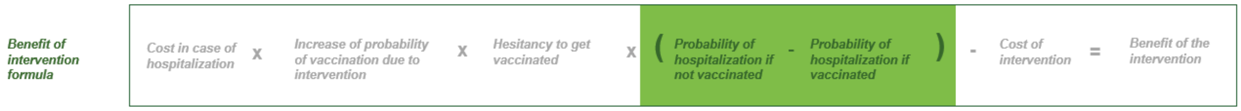
EXHIBIT 6-1



ID	intervention	hesitancy	cost_of_risk	cost	int_effectiveness
1MObcfaSTac85Lca0Y8bbA6l	1. Digital reminder	0.620485	20470	0.01	0.03
1MObcfaSTac85Lca0Y8bbA6l	2. Newsletter	0.620485	20470	0.02	0.14
1MObcfaSTac85Lca0Y8bbA6l	3. Phone call	0.620485	20470	20.00	0.18
1MObcfaSTac85Lca0Y8bbA6l	4. Discount	0.620485	20470	50.00	0.30
1MObcfaSTac85Lca0Y8bbA6l	5. Cash disbursement	0.620485	20470	50.00	0.30
5M89OSTL580dYeA849d3480l	1. Digital reminder	0.678411	20470	0.01	0.03
5M89OSTL580dYeA849d3480l	2. Newsletter	0.678411	20470	0.02	0.14
5M89OSTL580dYeA849d3480l	3. Phone call	0.678411	20470	20.00	0.18
5M89OSTL580dYeA849d3480l	4. Discount	0.678411	20470	50.00	0.30
5M89OSTL580dYeA849d3480l	5. Cash disbursement	0.678411	20470	50.00	0.30

The only missing ingredient to compute the benefit of the intervention is the % of hospitalizations for patients vaccinated and not vaccinated, i.e. the highlighted terms in the following exhibit.

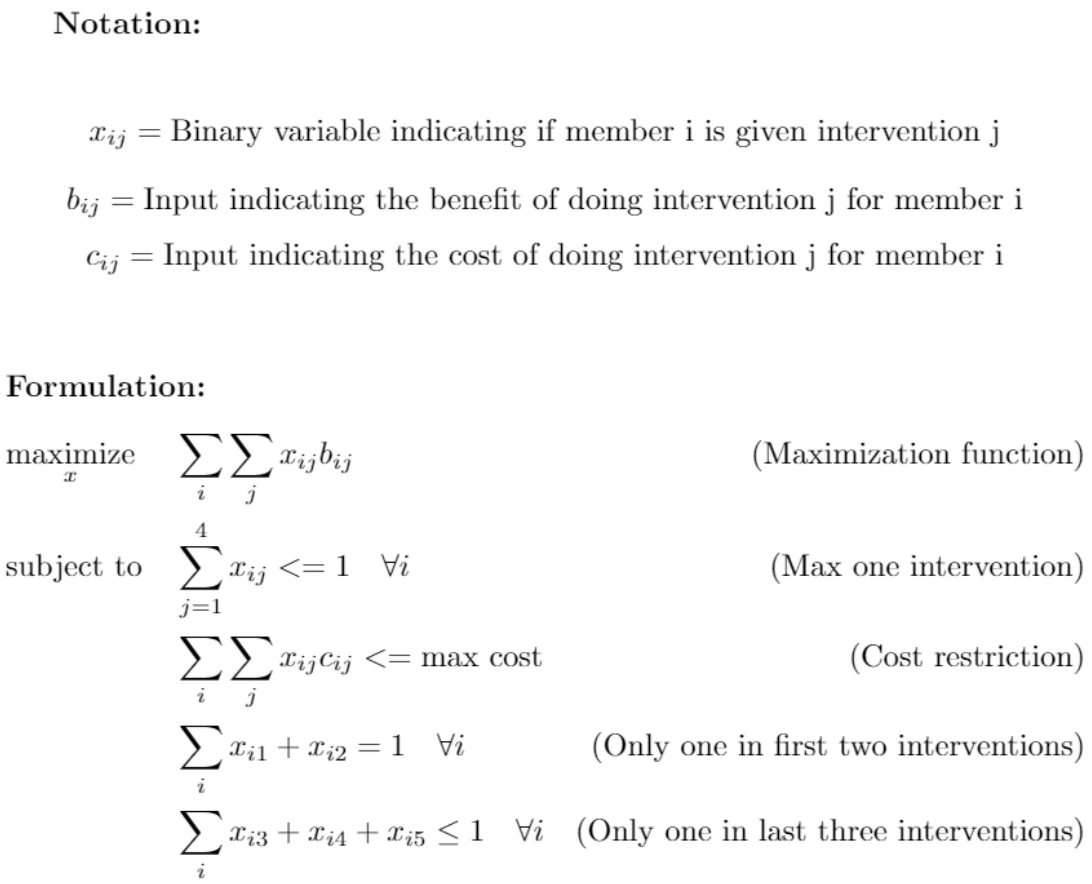
EXHIBIT 6-2


Benefit of intervention formula
Cost in case of hospitalization
x
Increase of probability of vaccination due to intervention
x
Hesitancy to get vaccinated
x
(
Probability of hospitalization if not vaccinated
-
Probability of hospitalization if vaccinated
)
-
Cost of intervention
=
Benefit of the intervention

To this end, we have leveraged data from effectiveness (29-fold reduction in hospitalizations) and from monthly hospitalizations[33] (360k at the time of the data gathering in March for the population of interest) to estimate the probabilities of hospitalization if not vaccinated and vaccinated at 0.8% and 0.03% respectively. With this last datapoint in hand, we compute the benefit of intervention for each pair of Humana member-intervention. Having the pairs member-intervention, benefit, and cost, we have formulated an optimization problem to maximize the captured benefit with the restriction of not exceeding a given cost and not assigning more than one intervention to the same member as defined in the next page.

[33] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html

EXHIBIT 6-3



Notation:

$$x_{ij} = \text{Binary variable indicating if member i is given intervention j}$$

$$b_{ij} = \text{Input indicating the benefit of doing intervention j for member i}$$

$$c_{ij} = \text{Input indicating the cost of doing intervention j for member i}$$

Formulation:

$$\begin{aligned}
\underset{x}{\text{maximize}} \quad & \sum_i \sum_j x_{ij} b_{ij} && \text{(Maximization function)} \\
\text{subject to} \quad & \sum_{j=1}^{4} x_{ij} <= 1 \quad \forall i && \text{(Max one intervention)} \\
& \sum_i \sum_j x_{ij} c_{ij} <= \text{max cost} && \text{(Cost restriction)} \\
& \sum_i x_{i1} + x_{i2} = 1 \quad \forall i && \text{(Only one in first two interventions)} \\
& \sum_i x_{i3} + x_{i4} + x_{i5} \leq 1 \quad \forall i && \text{(Only one in last three interventions)}
\end{aligned}$$

Where we basically want to find the values of x_ij which denote the choice of selecting intervention j for member I with the cost restrictions and the restrictions of only being able to choose either newsletter or digital reminder and one of the last 3 interventions at the same time.

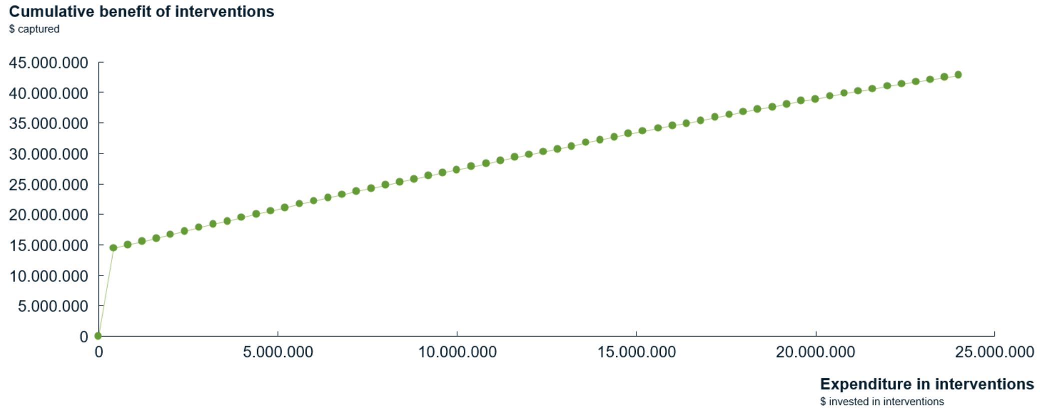
Once the optimization problem has been formulated as above, we have solved it with the open-source optimization package Pulp[34] to find the results that will be analyzed in the following subsections.

[34] https://github.com/coin-or/pulp

6.1 OPTIMAL ASSIGNMENT

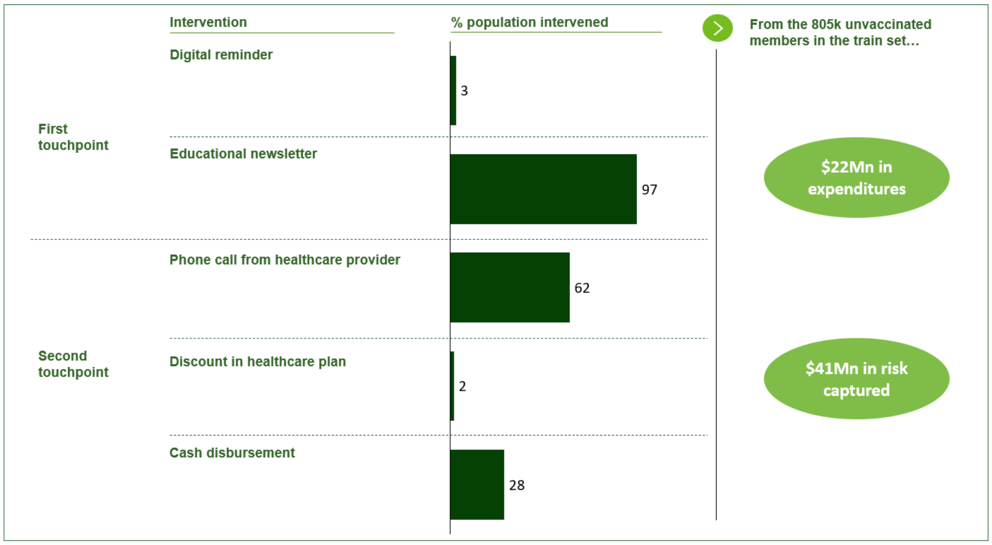
After executing the optimizer at different cost cutoff points, we retrieve the following graph which portrays the tradeoff between investing more capital against capturing benefits from the interventions. Each point has been calculated executing the optimizer described in the previous section with a different cost cutoff for the 805k members not vaccinated from the train set.

EXHIBIT 6-4



As one can note, from a certain point onwards the relative returns are diminishing yielding that the optimal breakeven point is at $22Mn. Setting this as our plan, $41Mn in risk revenue could be captured which would total a $19Mn in risk profit captured - we can double click to analyze which interventions we would actually be performing. In exhibit 6-5 we see that the first touch point would be mostly newsletter and that 8% of the members would not be intervened in the second touchpoint.

EXHIBIT 6-5



For all, we can conclude that our effort to quantify the benefit of each combination of member-intervention has yielded an optimized list of interventions to conduct that would capture an estimated $XXXMn at a cost of $XXXMn.

7 Appendix

This seventh and last chapter of the study conveys very relevant content that did not find its way in the natural storyline of the previous chapters. From potential enhancements to some notes on the assumptions made to finally a reference to the codebase and bibliographic literature leveraged.

7.1 APPROACH ASSUMPTIONS

Throughout the development of the study different assumptions have been made to quantify the different elements of the benefit equation as below:

EXHIBIT 7-1

Benefit of intervention formula

Cost in case of hospitalization x *Increase of probability of vaccination due to intervention* x *Hesitancy to get vaccinated* x (*Probability of hospitalization if not vaccinated* - *Probability of hospitalization if vaccinated*) - *Cost of intervention* = *Benefit of the intervention*

While hesitancy to get vaccinated comes directly from the proprietary Humana data, the other components have been estimated in an outside-in fashion and could be further refined – to reflect the present state, to include the particularities of Humana’s context and fine-grain them to be personalized, more precisely:

- **Linearity of contribution of effectiveness**: While there might be many possible relations in which interventions affect the decrease of hesitancy, as a simplification we have assumed that it is linear with constant factors for each different intervention
- **External validity of the effectiveness estimates**: While the effectiveness estimates of the different interventions have been pulled from other studies, they might not apply to this particular context or could be fine-grained. In order to have a more accurate proxy of these it would be necessary to conduct a randomized control experiment to see how different subpopulations would react to the different interventions
- **Only the costs come from hospitalizations**: We are simplifying the approach to only consider hospitalization costs attributable to the health insurer – network effects, long term health loss and other factors are not accounted for
- **There are no restrictions to reach out to the population**: An underlying hypothesis throughout the study is that there is freedom to reach out and intervene on every possible member of *Humana*, this might not be possible since there could be limits to the number of times they are reached out, *Humana* does not have their contact information …
- **Fine-tune the cost of risk for *Humana* for each member**: In this study we have mapped average hospitalization costs for different subpopulations – assigning an individualized cost of risk could improve the faithfulness of the final recommendation

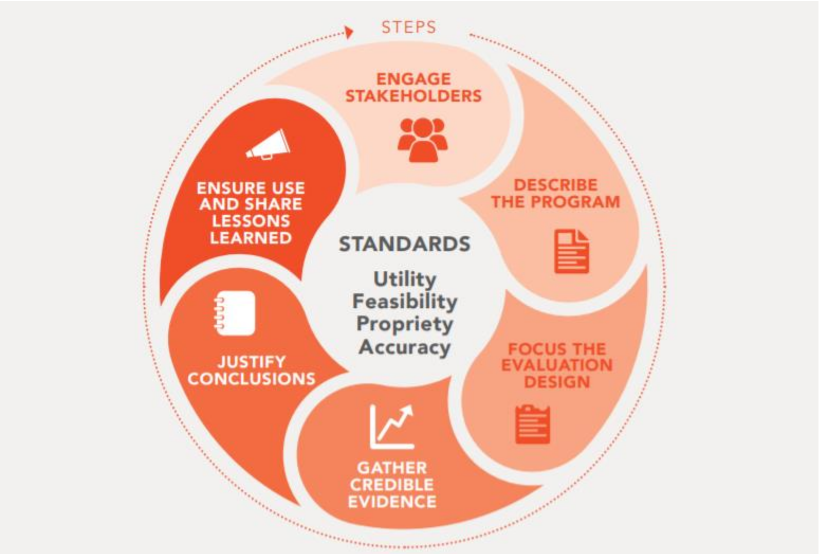
7.2 POTENTIAL ENHANCEMENTS

Besides refining the different elements from the benefit equation as describe din 7.1, in this subsection we itemize alternative possibilities to improve the results of the study by enlarging its scope:

- **Consider a rolling time period**: Instead of having a fixed snapshot of the data and behavior of the members, a fixed time frame to train and to predict with
- **Consider more interventions**: Include all the array of potential interventions that *Humana* can deploy
- **Relate the causes of not being vaccinated with the interventions**: In this study we have treated the predictive task as an error-minimizing exercise but have not built the causal graph that would explain the reasoning behind not getting the vaccine. This graph could be created with field experts and additional data and would allow to map interventions to causes of hesitancy.

Furthermore, as outlined in the WHO vaccination intervention guideline[35], it does not suffice to execute interventions (steps 1-8 of their guideline) but a crucial feature of the success of any roll out is its continuous monitoring, in the following exhibit one can observe the steps that the WHO recommends following once a plan of our characteristics is launched.

EXHIBIT 7-2



7.3 ACKNOWLEDGMENTS

As a closing note, we would like to place on record our deepest sense of gratitude to Humana and Mays Business School's departments involved with organizing, hosting, and facilitating the opportunity to delve into this fascinating dataset and allowing us to suggest solutions for problems with such societal relevance.

[35] https://www.who.int/immunization/programmes_systems/policies_strategies/MOV_Intervention_Guidebook.pdf