Estimating Migrant Population in Malaysia and Its Implications

Emir Izat Abdul Rashid[1*]

[1*]Center for Technology, Strategy, and Sustainability, Asia School of Business, Jalan Dato Onn, Kuala Lumpur, 50480, Kuala Lumpur, Malaysia.

Corresponding author(s). E-mail(s): emir.izat@asb.edu.my;

Abstract

How do we estimate the population of migrants in Malaysia? The uncertainty in estimating migrant population in Malaysia has posed public policy challenges in implementing suitable responses to address the vulnerabilities faced by them. And the variety of estimates by different scholars makes it hard to know which estimates are more reliable than others for policy-making. We would like to formalize assumptions that enable the comparisons between different estimates. We developed a model where a migrant with a legal status (either legal or undocumented) choose to vaccinate according to their preference and probability of being caught and getting infected with COVID-19. Then, from the assumption of the parameter values of the model and the distribution of the COVID-19 infection function, we can derive the estimate of undocumented migrants in Malaysia. Our benchmark estimate is consistent with the lower bounds typically publish by NGO and international organizations $\sim$ 2 million undocumented migrants. Changes in the parameter values and assumptions allow us to quantify and formalize the implications of empirical estimates made by other organizations and scholarly work.

Keywords: migration, undocumented migrant, credible estimation, COVID-19

JEL Classification: J510 , J610 , J180 , I1

1 Introduction

A prominent issue arise for a wide range of social science research require the examination of phenomena undefined during data production. For example, during census enumeration, undocumented migrants may not be defined as such because the option is not available during the counting process. In most cases, social science researchers ranging from demographers, sociologists to a wide variety of other fields have formulated sampling techniques that accommodate definitions undefined during the production of data. However, the wide variety of assumptions that each technique use to define the sample, to collect them in the field, to process them into usable datasets, and to analyze them has led to potentially non-replicable findings. Yet, despite the mixed findings of different studies, how can we formulate a technique to compare and contrast different estimates or findings? What are the methods where given two population estimates – previously undefined by verifiable sources – using different methods, we can formulate basic assumptions in which we can identify and quantify the source of bias?

This question is an important one in the social sciences due to the replication crisis that took place in the early 2010s, where many earlier studies – including highly established ones – are found to not be able to stand the test of time with modern instruments and data Nosek et al. (2022). This includes estimates of population sizes especially hidden populations such as undocumented migrants. On the one hand, researchers can conduct their own survey of a random sample. However, not many are quick to admit their legal status out of fear of being caught and face consequences. A classic example of the weakness of these types of surveys is electoral polling; there has been a political science literature on the hesitance of Republicans to respond sincerely which led to unanticipated outcomes in 2016 Guess et al. (2019). On the other hand, researchers can perform ”respondent-driven” surveys where they leverage on network theory to find their data points and estimate the quantity of interest Heckathorn (1997) Naderifar et al. (2017). As a result, there has been a proliferation of studies where marginalized communities are the main subject of research: leveraging on the community network enables researchers to estimate different parameters of a community that are hidden in plain sight. There has been work to estimate population parameters of a diverse set of communities from HIV/AIDS patients to small Native American tribes Johnston et al. (2016) Magnani et al. (2005).

We would like to reconcile these two approaches in estimating the population size using a sampling method that is in between researcher-driven and respondent-driven: public health government program, specifically the distribution of vaccines. Vaccine distribution or more generally public health is one of the most important consideration in both individual's life and the government. Individuals need public health to maintain a high quality of life while governments need public health to maintain a healthy population. Thus, in public health emergencies such as COVID-19, governments are willing to over-supply vaccine to ensure every individuals receive them – and this include individuals who are not legally defined due to public health concerns outweighing national security. Individuals are motivated to get vaccinated to avoid the adverse symptoms of COVID-19. Therefore, this aligning of interests enables us to estimate the population where governments supply vaccines at their vaccine centers (researcher-driven sampling) and individuals choose to vaccinate (respondent-driven sampling). This would allow us to compute the population of the migrant in a much more robust way, whilst clarifying the assumptions to compare different population estimates from multiple sources for example, the United Nations High Commissioner for Refugees, the World Bank, NGO, etc.

To formulate this relationship more precisely, I used a simple economic model where I assume supply of COVID-19 vaccine is unlimited[1]. Then, I developed a model where individuals either legal or undocumented choose to vaccinate depending on their preference to vaccinate (whether they gain or lose from vaccinations), their probability of being caught by the police, the punishment of being caught, and the severity of COVID-19. Then, optimizing their utility function, individuals will choose to vaccinate depending on the parameter values.

To illustrate the model's ability to clarify assumptions, I developed a benchmark estimate where I assumed specific parameter values in the model. The benchmark estimate that I computed resulted in approximately 2 million undocumented migrants currently residing in Malaysia. I then performed a comparative statics to see the conditions and assumptions behind potential candidates of estimates. I established that the lower bound for undocumented migrants in Malaysia is 1.2 million but this could only be achieved through the model's process if either those who refuse to vaccinate did not lose utility from getting vaccinated or every migrant prefers vaccine – this estimate is more aligned with the World Bank's estimates Yi et al. (2020). On the other side of the spectrum, the upper bound of the population estimate (8.24 million migrants) Nixon (2024) would either imply either 75% migrants are deterred from receiving vaccines due to fears of punishment or either 80% of migrants do not prefer vaccines at all.

The simple economic model illustrates the effect of basic assumptions can have towards empirical models and observations. In section 2, we will briefly describe the context and the literature review; in section 3, we will be formulating the model of migrant's vaccination choice; in section 4, we will be presenting the assumptions we make in the benchmark model; in section 5, we will present the empirical findings of our benchmark model; in section 6, we will be discussing the potential implications to empirical research.

2 Context and Literature

2.1 Institutional Background

Human migration have long been part of the societies in the Malay Archipelago as a result of trade resulting from its geographical location connecting China to Europe. Starting from the Indian Ocean trade, human migration is the cause for the spread of Islam in the region Gungwu (1985). Malaysia in particular, human migration in the late 19th century to the early 20th century saw a rapid inflow of migrants from China and India as a result of economic opportunities provided by the growing world demand for tin and rubber as well as serving its traditional role in trade. After gaining

[1] This assumption may be very difficult to achieve in low-income countries but in the case of Malaysia, there is an extra 8.5 million unused COVID-19 vaccine doses.

independence from the British, the next wave of migration began in the 1980s as a result of industrialization and economic growth. Malaysia today remains a migration hub as it was in the past.

The most common nationalities that make up the foreign migrant population today comes from Indonesia, Bangladesh, Nepal, Myanmar, and India for Peninsular Malaysia while the Philippines provides the largest portion of foreign migrants in the state of Sabah Mobarak et al. (2023) Anderson (2021). Most of these migrants arrive to the country during the times of rapid economic growth in the 1980s – present. The government tightly controls the nationality and the types of work allocated to them. For example, Nepalese migrants can only work exclusively in security and no other sector of Home Affairs. Migrants are granted work passes and regularly require an agent to travel Anderson (2021). However, whether this regulation is tightly complied by both migrants and employers remain an open question.

The reason why foreign migrants present a challenge in government is because of a lack of legal protection afforded by the state especially the undocumented ones Pandi et al.. Since the undocumented is not recognized by the government, they bear no obligations to include them in policy-making. As a result, the undocumented migrants are also one of the most vulnerable groups in Malaysia due to the larger role of discretionary power by the state and other organizations when they have to make decisions on their behalf Secretariat Devadason and Meng (2014) Theng et al. (2020). For example, if an undocumented migrant wanted to work, employers bear the risk of hiring them thus are more prone to underpay them as a way to compensate the potential losses Pandi et al.. Not only that, the government typically conducts raids where both migrants, documented and undocumented, are questioned and investigated Putul and Mia (2018). The uncertainty in terms of where and when the government will take action creates hostile living conditions for migrants – documented and undocumented alike. This environment does not allow for other government agencies such as the census to count them as they may not be as enthusiastic to respond to census or survey enumerators.

Yet, neglecting to account foreign migrants in policy-making has caused enormous losses for the government. In public health, due to the barriers to home rentals and ownership, migrants are more likely to live in cramped spaces where disease such as influenza and tuberculosis thrive Theng et al. (2020). And the living situation of migrants presents an overall cost to everyone because diseases spread without the discrimination of documentation. Not only that, improperly accounting for migrant populations have major implications to the labor market – a source of potential backlash for many native population Abramitzky et al. (2014) Tabellini (2020) Guriev and Papaioannou (2022).

As a result of the COVID-19 pandemic, the government launched one of the largest vaccine drives the country has seen in its history. Due to concerns of public health outweighing the concerns of national security, the government has promised to not take legal action against undocumented migrants who wish to vaccinate. The outcome of the COVID-19 vaccine drive is the release of one of the most comprehensive publicly-available population dataset sort by district, nationality, ethnicity, gender, age group, and comorbidity. We will be using this dataset in the subsequent section.

Another benefit that we have in using the vaccination data is the fact that the border during this period of time is closed which means that the flow of migrants is likely to be small and the population is more or less "static". Although one could argue that the estimation does not take into account current population numbers nor population flows, I argue that estimating the migrant population between the beginning of COVID-19 towards the border opening is important as these are the types of people who are not as mobile, those who are less likely to move out which means the government should account them in the policy calculation.

2.2 Literature Review

The most extensive review of migrants and the procedures of arrival and exit was conducted by Yi et al. (2020). In that paper, they attempted to estimate the population of foreign migrants through the combination of many different datasets on top of the 2010 census such as entry pass data, remittance data, and medical screening tests for work among others. Their empirical estimates suggests that between 1.2 to 1.5 million undocumented migrants. Hwok-Aun and Leng (2018) arrived at a conservative estimate of migrant population to be around 4 million people and can be as high as 5.5 million which puts the undocumented population as high as 3.7 million. A review of all the studies on population estimation ranges from as low as 1.2 million and as high as 8.24 million people Nixon (2024). How do we make sense of the myriad of all these estimates? Although most of the studies in this strand attempt to use copious and careful use of a universe of available datasets, we would like to develop a framework in which we can take a population estimate and quantify the implications of the estimate to infer the composition and the choices that migrants make to appear as a data point. Thus, we seek to add onto the literature of population estimation by utilizing economic methods and blending it with empirical observations through accounting for self-selection.

The framework we use can best be described in line with the emerging literature of theoretical implications to empirical models Slough and Tyson (2023) Bueno De Mesquita and Tyson (2020) Izzo et al. (2018). Although this tradition is rooted in using theoretical models to develop and evaluate empirical designs that studies causality between two or more variables, theoretical reasoning can also be used to evaluate population estimate with relatively simple economic tools. Thus, we seek to add on the literature through an application of this framework in the domain of demography.

Not only that, we seek to add on the literature of estimating hidden populations especially by developing a framework that enables a researcher who use different types of sampling strategy either researcher-driven or respondent driven to quantify the selection bias inherent in their sampling design Johnston et al. (2016) Léon et al. (2016) Gile and Handcock (2010). Although our question is narrow and may not apply to other contexts of hidden population, the reasoning employed can prove to be useful to clarify the findings made about the population studied.

2.3 Data

The data used is the COVID-19 vaccination data provided by the Ministry of Health. The data includes the number of daily COVID-19 vaccinations, the district in which they are given, the state, and the nationality of the people that receive them. The data is then summarized to find the total number of people (Malaysian or not) that vaccinated between January 2021 – April 2022. The reason why I pick that date is because April 2022 is the border opening which would require one to take into account migrant inflows. As for the population estimates, I have used the census data provided by the Department of Statistics.

3 Theoretical Model

The model is a constrained optimization where an individual with legal status $\tau = \{l, u\}$ and vaccine preference $\nu = \{v, av\}$ have to choose whether to vaccine. Legal status is either legal l or migrant u. Vaccine preference is whether they are comfortable with receiving vaccine. If $\tau = u$, vaccination carries a risk of getting caught $Pr(\text{caught}) = p$ and getting caught incurs a cost c where both values are exogenous. If $\nu = av$, vaccination reduces utility by r_{av} and if $\nu = v$, not vaccinating reduces the utility by r_v. If individual do not vaccinate, there will be an idiosyncratic shock $\eta_i \sim \mathcal{F}$ where $\mathcal{F}$ is a CDF with support ranging from $(0, +\infty)$ that conveys the level of severity of COVID-19.

Utility function for a legal migrant is:

$$U_l = \begin{cases} -\mathbb{1}_{\nu=av} r_{av} & \text{If vaccinate} \\ -\mathbb{1}_{\nu=v} r_v - \eta_i & \text{If not vaccinate} \end{cases}$$

Utility function for an undocumented migrant is:

$$U_u = \begin{cases} -\mathbb{1}_{\nu=av} r_{av} - pc & \text{If vaccinate} \\ -\mathbb{1}_{\nu=v} r_v - \eta_i & \text{If not vaccinate} \end{cases}$$

The outcome of the model is as such:

- If $\tau = l, \nu = v$, probability of vaccination is 1.
- If $\tau = u, \nu = v$, probability of vaccination is $Pr(\eta_i \geq pc - r_v)$
- If $\tau = l, \nu = av$, probability of vaccination is $Pr(\eta_i \geq r_{av})$
- If $\tau = u, \nu = av$, probability of vaccination is $Pr(\eta_i \geq pc + r_{av})$.

Thus, how many people will vaccinate? Suppose that $N_{\tau\nu}$ is the population for τ, ν:

$$N_{\text{vaccinate}} = N_{lv} + Pr(\eta_i \geq pc - r_v)N_{uv} + \\ Pr(\eta_i \geq r_{av})N_{lav} + Pr(\eta_i \geq pc + r_{av})N_{uav}$$

$$N_{\text{no vaccinate}} = 0 + F(pc - r_v)N_{uv} + \\ F(r_{av})N_{lav} + F(pc + r_{av})N_{uav}$$

From this framework, we will gather the assumptions to give substantive meaning to an estimation figure by a researcher, policymaker, or any other organization.

4 Assumptions of Parameter Values

Before using the data to estimate migrant population, we will first clarify a few basic assumptions that will form our benchmark case. Although we know that these assumptions may not be true in reality, we will show how relaxing these assumptions can lead to starkly different population estimates. As we will show in section 5, we will be using this tool of breaking one assumption while maintaining others to show which parameters will contribute towards overestimation and which parameters will contribute towards underestimation.

4.1 Benchmark Assumptions

The benchmark assumptions we need to make are first and foremost that vaccine penalty $r_{av} \to \infty$. This assumption implies that for any parameter values, vaccine preference $\nu = av$ will not vaccinate because the probability to vaccinate will approach zero. No matter how severe the COVID-19 may be, these individuals will not vaccinate. The real-world interpretation of the individuals of this type are the individuals who will not vaccinate no matter how much facilities are provided to them to vaccinate that is the never-takers. The population that could be included in this group are the vaccine skeptics, small children who are not qualified to receive them, and individuals who do not vaccinate due to social norms stemming from religious beliefs.

Then, we also assume that probability of being caught and cost of punishment $p = 0, c > 0$. The reason why this assumption may be justifiable is due to the fact that the government have committed to not take legal action against migrants. This assumption also allows us to say that although the cost of being caught is non-zero that is undocumented migrants will face some consequences for being caught, the probability of being caught by authorities is 0 which means that every vaccine preference $\nu = v$, legal or undocumented will vaccinate for all vaccine reward $r_v \geq 0$.

Finally, we also will have to assume that the vaccine preference ν, legal status τ, and nationality are uncorrelated. This assumption would be necessary for us to generate the distribution of people who are $\nu = v$ and $\nu = av$. Although it is clear that there will be some unobserved variation in cultural or social aspects of vaccine preference, we will assume that the proportion of people who are $\nu = v$ or $\nu = av$ is the same regardless of nationality and legal status.

Since we know that all three assumptions above are met, we can use Malaysian nationals to compute the expected proportion of vaccine preferences ν for l, u, which will all equal each other. Then, using this expected proportion combined with the population all legal migrant, we derive the population of legal and undocumented migrant who are vaccine-friendly $\nu = v$. We derive the population of migrant who are vaccine-averse $\nu = av$ which enables us to derive an estimate of population of undocumented migrant.

Population	$\nu = v$	$\nu = av$	Total population
Malaysians	24,096,130	6,102,070	30,198,200
Documented Migrants	1,994,991	505,209	2,500,200
Undocumented Migrants	1,728,745	349,323	2,077,068
Total Migrants	3,723,736	854,533	4,578,269

Table 1 Migrant Estimates

Note: Red color indicates a value that is derived.

5 Empirical Results and Comparative Statics

5.1 Migrant Estimation of Benchmark Assumption

For the census data, we obtained that the population size of Malaysians is 30,198,200 people while documented migrants is 2,500,200 people. From the Ministry of health, we obtained that 24,096,130 Malaysians and 3,723,736 non-Malaysians vaccinate – which means that about 20% of Malaysians are not vaccinated. We use this proportion and the uncorrelated assumption as well as the vaccine penalty $r_{av} \to +\infty$ assumption to estimate the population of vaccine preference $\nu = v, \nu = av$ for the documented migrants. Then, from the estimate of documented population and the probability of being caught $p = 0$ assumption, we obtain the population estimate for the undocumented migrants for $\nu = v$. From the vaccine penalty $r_{av} \to \infty$ we manage to estimate the undocumented migrant population who are vaccine-averse $\nu = av$. Our benchmark estimate suggest that there is 2,077,068 migrants between 2021 – 2022.

5.2 Comparative Statics

Now, what happens when we relax the assumptions in the benchmark? First, suppose that vaccine penalty r_{av} decreases that is all else equal, the reduction of utility penalty faced by the vaccine-averse $\nu = av$ when they vaccinate. When this happens, the probability of vaccination for the vaccine-averse $\nu = av$ increase, so the population composition of those who are vaccinated will include $\nu = av$ and $\nu = v$. When vaccine penalty r_{av} goes to 0, every migrant, legal and undocumented, will be vaccinated. In this case, the estimated undocumented migrant population will be 1.22 million. Thus, reducing r_{av} will reduce the population estimate of undocumented migrants with a baseline of 1.22 million.

Assume now as the benchmark case but probability of being caught $p \neq 0, c, r_v$ such that $pc - r_v > 0$. We know that $\nu = av$ does not vaccinate. Probability of vaccination $Pr(\eta_i \geq pc - r_v) \neq 1$ which means undocumented population estimate will be higher than benchmark. Thus, increasing p will always increase the estimated population of the undocumented migrants.

Assume again the benchmark case but the composition of vaccine preference ν of migrants is not equal to the Malaysian nationals case. If the portion increase say from 80% to 95%, then the population of undocumented migrants decrease from the benchmark case. If the portion decrease say from 80% to 65%, then the population of undocumented migrants increase from the benchmark case. To interpret this in the real world, composition effects are driven by the correlation between the identity of the person and the preferences they have. For example, certain cultures may be more averse to vaccines which may not be true but the model accommodates for such explanations. Thus, changing population mix in terms of vaccine preference can also drive estimation biases where increasing the composition of $\nu = v$ undocumented population decrease and decreasing the composition of $\nu = v$ increase the population of undocumented migrants.

5.3 Assessment of Population Estimates

Using the framework we have explored above, we can now begin to assess the implications of the estimates provided by external organizations. For this section, we will be analyzing the most conservative estimate which is one put forth by the World Bank, a middle estimate which is one put forth by Hwok-Aun and Leng (2018), and a liberal estimate which is one put forth by Nixon (2024). Although their estimated population is for years prior to the study around $2017 - 2018$, we do not expect to find drastic population shifts to change in a 5-year period even after accounting for COVID-19 due to lockdowns that is no sudden shocks in people leaving. Thus, we will take their estimate as given for the year $2021 - 2022$.

For the World Bank estimate, their population estimate is around 1.23 to 1.46 million. Our model predicts two effects: reduction in vaccine penalty r_{av} and composition effects. As mentioned in 5.2, our baseline estimate if $r_{av} = 0$ will mean that the population of undocumented migrants would yield a population of 1.22 million. Thus, if we only account for r_{av}, the slight increase in their estimate would mean r_{av} is close to 0. To interpret this, conditional on being a migrant, those who are averse to vaccines are not averse enough to deter from the consequences of not vaccinating. If we now assume that $r_{av} \to \infty$, the other way that this estimate can be achieved is we also find that migrants are more vaccine friendly that is higher portion of vaccine-friendly $\nu = v$. The model will predict that the portion of $\nu = v$ among migrants is 94%. We showed here that the World Bank strongly relies on a high degree of vaccine compliance from foreign migrants which can be a difficult implication to accept. The main weakness of this estimate is the use of administrative data for computing the estimate. Since undocumented migrants tend to not have any records in the first place, they acknowledged the under-estimation problem in their paper. We quantified what this under-estimation should imply in observed vaccination trends for migrants.

Nixon (2024) estimated undocumented individuals implied through rice consumption, which he estimated to be $6.28 - 8.24$ million. Our model predicts two effects: probability being caught $p \neq 0$ or composition effects. If we only account for $p \neq 0$ without composition effects, the 8.24 million estimate would imply that 75% of migrants are deterred from getting vaccines. If we only account for composition effects, the 8.24 million estimate implies that 80% of migrants do not prefer vaccines at all which means that there is a widespread aversion towards vaccines among the migrant population. Of course, both effects can be combined to produce the estimate as proposed by Nixon but empirical evidence of what the parameter values are is needed for the validity of the estimate. In contrast to the World Bank, Nixon's estimate implies a high degree of vaccine non-compliance among the foreign migrants, which is a fair implication to accept but could be regarded as too excessive to accept.

For the Hwok-Aun and Leng (2018) estimate of 2.1 million – 3.7 million population, their lower bound corresponds nicely with our benchmark estimate. As for their upper bound, our model predicts that two effects: probability of being caught $p \neq 0$ or composition effects. If we only account for $p \neq 0$ without composition effects, 3.7 million undocumented migrants would imply that the probability of vaccination for the undocumented given they are vaccine-friendly is 46% which means that 53% of undocumented migrants are expected to be deterred from vaccine purely from fear of getting caught. If we assume that the authorities' claim is credible – $p = 0$ – composition effects would predict that 60% of migrant is $\nu = v$ which means there is a sizable minority averse towards vaccines among the migrant community compared to the natives. Of course, some combination of both effects can produce estimates within the proposed range. Hwok-Aun and Leng (2018) estimate is the most plausible out of the three because the implications are not too restrictive as the World Bank but not too excessive as Nixon's.

Overall, the model not only allows us to compare and contrast different estimates but to quantify them to see which parameter values may be the most plausible to accept. Although we are not adding anything in terms of better estimation techniques for a more accurate count, it establishes a framework in which one can assess the validity of one's population estimate. However, the model also enables us to see which parameters to control when we perform empirical tests with smaller samples to estimate the model parameters which we can then use to estimate the population of undocumented migrants.

6 Implications to Empirical Research

Thus, what is the future direction to better improvements in estimating the migrant population in Malaysia? A natural follow-up to the theory developed above is to connect the parameter in the model to real-world data and using this insight to develop empirical tests to see if an estimate is an under-estimation or over-estimation. And this correspondence between theoretical models and empirical observations enables us to compare different sets of estimation which can mean different policy implications.

6.1 Example of an Empirical Test to Estimate Population

To develop an empirical test to find parameter values, we need to maintain some of our benchmark assumptions to gather information of other parameters for the estimation of population. Ideally, we would do a random sampling where we ask each person for their identification and legal status and then compute their relative weights to the population of the country. However, as explained in section 1, undocumented migrants are less visible in census enumeration which pose a self-selection issue. Plus, ethical issues may arise such as potential security concerns if we make undocumented migrants a subject of our study.

From our benchmark assumptions, the most likely assumption to hold is vaccine penalty $r_{av} \to \infty$. This is because this assumption relies on our definition of what vaccine preference $\nu = av$ is. To simplify our empirical test we can define $\nu = av$ as individuals who would not vaccinate under any circumstance. Next, another plausible assumption which is likely to hold is the no correlation between vaccine preference ν and legal status τ. This assumption may hold because there is a lot of overlap of nationalities between documented and undocumented migrants. We do not expect that there is a stark cultural or social norm divergence between documented and undocumented migrants in terms of their preferences of vaccines. Another assumption we make is that there is no sharp inflow or outflow of migrants that is there exist no sudden shocks in population trends. Last but not least, we have to assume that the

k	q	N_u	Deterred from Vaccinating
0.90	0.8	2,045,606	368,389
0.65	0.5	6,457,249	2,098,606
0.50	0.9	5,496,969	274,848
0.95	0.5	2,839,044	1,348,546

Table 2 Empirical Test of Different Parameters

probability of getting caught is $p = 0$ or at least close to 0; this assumption is the weakest to make and we can relax this assumption to predict the number of people that are deterred from receiving vaccines. Thus, these four assumptions are sufficient to conduct an empirical test to estimate the population of undocumented migrants.

An example of an empirical test is a random sample survey where we ask migrants of any legal status of their vaccine preference ν. Using our definition of no vaccination under any circumstance, we can find in our sample the composition of migrants who will not vaccinate under any circumstance and the migrants who will vaccinate. Then, assuming that we conducted our survey to a level of satisfactory standards, we compute the composition of vaccine preference in the population. Suppose we find that the percentage of $\nu = v$ is k. We could now compute the population of undocumented migrants with these steps. Given we know that the population of documented migrant is N_l and label $N_{\tau\nu}$ as population of migrant of legal status τ and vaccine preference ν. N_v is the vaccinated population.

1. $N_{uv} = N_v - k(N_{lv})$
2. $N_u = k(N_{uv})$

If we decide to relax the probability of being caught assumption to $p \neq 0$ and determine the probability of vaccination for $\nu = v, \tau = u$ as q. Of course, we take this probability as exogenous. A diligent researcher may estimate this by identifying the people who prefer vaccines and whether they end up vaccinating. This estimate will be difficult to elicit due to potential social desirability bias. Then, the undocumented population will be:

1. $N_{uv} = (1 - q)(N_v - k(N_{lv}))$
2. $N_u = k(N_{uv})$

To give some concrete predictions here is an example of an output of an empirical test in table 2.

As noted in the table, the implication of this empirical test is also the identification of people who are deterred from vaccinations which is a policy-relevant quantity. Notice here that different policy environments require different policy implications. In the third estimate, there is a lot of undocumented migrants but high compliance which means not a lot of individuals being deterred from vaccinating but a high $\nu = av$ population – policy advice here is to reduce the value of r_{av} perhaps through monetary transfers to increase vaccine uptake. In contrast, the fourth estimate has less undocumented migrants but low compliance which means a lot of individuals being deterred from vaccinating – policy advice here is to reduce the cost of punishment c or reduce the probability of being caught p to increase vaccine uptake. Notice here that different parameter values lead to different policy conclusions and valuation of costs. On the one hand, monetary transfers is less costly and easy to implement whereas on the other hand, reforming the police department to be more lenient is a lot costlier politically.

One can relax more assumptions to obtain a more realistic figure but a more flexible design is prone to social desirability bias and potential ethical issues due to more information being requested from a marginalized social group. Thus, there is an inherent trade-off in information – more information can lead to high variation in estimates due to bias but less restrictive assumptions are needed. Different empirical specifications based on the model would need the modification of the assumptions stated in the benchmark model.

6.2 Model Weakness and Possible Extensions

As for the model specifically, there are many improvements we can make to sharpen the assumptions we make. First, we assume there is only two types of migrants $\nu = \{v, av\}$. A much richer model can incorporate heterogeneity in vaccine preference on a continuous line where one can have a ν ranging from $[0, 1]$ where 0 indicates no vaccines under any circumstance and 1 indicates taking vaccines under any circumstance.

Another assumption we do not take into account is time variation in the utility penalty of r_{av}. The COVID-19 is assumed to be the same throughout the period of $2021 - 2022$. We know that there are periods where demands of vaccine is higher than others; for example, during periods when COVID-19 spikes, r_{av} may decrease and during period when COVID-19 has stagnated r_{av} may increase.

We also do not take into account the heterogeneity in geography. The reason why this may play a role is because cost of vaccination may vary depending on the distance between the individual and the vaccine center or clinic. Thus, migrants living in the city will be more likely to vaccinate than migrants who live in rural areas. This may pose a problem if there is a bias on the types of migrants who live in either rural or urban areas. For example, undocumented migrants may prefer to live in rural areas because they can avoid authorities better there in urban ones. However, one can also argue that this motive may be offsetted by economic reasons as urban areas have better prospects for social mobility Abramitzky et al. (2014) Borjas (1987). Thus, a better model will incorporate this consideration and the data needed to show its implication.

Last but not least, our model will benefit a lot from considering multi-periods where inflow and outflow of migrants become important. A key ingredient in these types of models is the endogenous nature of migrant entering and exit decisions. While migrant is likely to leave and enter a country due to short-run economic considerations, equally important is the time they stay in their host country. Models of these types must take into account inter-temporal considerations which would further complicate the estimation of their population.

6.3 Conclusion

Aside from the geographic heterogeneity, there are many other covariates that can determine the probability of a migrant's choice in vaccination apart from the fear of getting caught. However, the simplistic nature of the model is to illustrate the power of theory in estimating the size of hidden population. The most salient feature of hidden population is self-selection. Any methodology that make marginalized communities a subject of their research depends a lot on the relationship of the researcher to the community that they study. Thus, the information they receive to conduct their research will be selective – a clarification of the researcher assumptions enables other researchers to compare different sets of findings. Although findings are no doubt important, equally important is the quantification of selection they are exposed to. I attempted to quantify this by examining the estimation of undocumented migrants in Malaysia.

One framework that can be used to compare different sets of undocumented migrant population estimate – which can be as low as 1.2 million and as high as 8.24 million – is the utility maximization framework. I then developed a constrained optimization model where migrants choose to vaccinate. The reason why I pick vaccination drives as a sample is because of the availability of data. I showed that if migrant's choice depends on their vaccine preference, probability of getting caught, and the severity of COVID-19 then assumptions of parameter values can yield population estimates. I then made benchmark assumptions to yield an estimate and showed that modifying these assumptions can yield wildly different population estimates. I used the model framework to assess different estimates from external sources such as the World Bank and other authors as well as showing which variables are the ones driving these estimates. I then used the model insights to build an empirical test that can be used to estimate the population of undocumented without directly asking them their documentation status.

We attempted to illustrate the power of theory in comparing different population estimates in much more robust and clear manner. The issue with population estimates is the different definitions of measurement. Theory clarifies these definitions in a standardized way so that we can start comparing different empirical results. Not only that, the theory we developed above also offer some insights into potential policy antidotes as a consequence of the model. As shown in section 6.1, the model enables us to identify the group of people who are prevented from receiving vaccines despite wanting them. Although better population estimation techniques are no doubt important for the advancement of policy or political debates, equally important is the theoretical implications of estimates which may offer competing interpretations.

References

Abramitzky, R., Boustan, L.P., Eriksson, K.: A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. Journal of Political Economy **122**(3), 467–506 (2014) https://doi.org/10.1086/675805 . Publisher: The University of Chicago Press. Accessed 2024-04-07

Anderson, J.T.: Managing labour migration in Malaysia: foreign workers and the challenges of 'control' beyond liberal democracies. Third World Quarterly **42**(1), 86–104 (2021) https://doi.org/10.1080/01436597.2020.1784003 . Publisher: Routledge _eprint: https://doi.org/10.1080/01436597.2020.1784003. Accessed 2024-04-07

Bueno De Mesquita, E., Tyson, S.A.: The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior. American Political Science Review **114**(2), 375–391 (2020) https://doi.org/10.1017/S0003055419000911 . Accessed 2024-03-11

Borjas, G.J.: Self-Selection and the Earnings of Immigrants. The American Economic Review **77**(4), 531–553 (1987). Publisher: American Economic Association. Accessed 2024-04-07

Devadason, E.S., Meng, C.W.: Policies and Laws Regulating Migrant Workers in Malaysia: A Critical Appraisal. Journal of Contemporary Asia **44**(1), 19–35 (2014) https://doi.org/10.1080/00472336.2013.826420 . Publisher: Routledge _eprint: https://doi.org/10.1080/00472336.2013.826420. Accessed 2024-04-07

Gile, K.J., Handcock, M.S.: Respondent-Driven Sampling: An Assessment of Current Methodology. Sociological methodology **40**(1), 285–327 (2010) https://doi.org/10.1111/j.1467-9531.2010.01223.x . Accessed 2024-03-11

Guess, A., Munger, K., Nagler, J., Tucker, J.: How Accurate Are Survey Responses on Social Media and Politics? Political Communication **36**(2), 241–258 (2019) https://doi.org/10.1080/10584609.2018.1504840 . Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2018.1504840. Accessed 2024-04-07

Guriev, S., Papaioannou, E.: The Political Economy of Populism. Journal of Economic Literature **60**(3), 753–832 (2022) https://doi.org/10.1257/jel.20201595 . Accessed 2024-04-07

Gungwu, W.: Migration Patterns in History: Malaysia and the Region. Journal of the Malaysian Branch of the Royal Asiatic Society **58**(1 (248)), 43–57 (1985). Publisher: Malaysian Branch of the Royal Asiatic Society. Accessed 2024-04-07

Hwok-Aun, L., Leng, K.Y.: Counting Migrant Workers in Malaysia: A Needlessly Persisting Conundrum (2018) (2018)

Heckathorn, D.D.: Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations*. Social Problems **44**(2), 174–199 (1997) https://doi.org/10.2307/3096941 . Accessed 2024-04-07

Izzo, F., Dewan, T., Wolton, S.: Cumulative Knowledge in the Social Sciences: The Case of Improving Voters' Information, Rochester, NY (2018). https://doi.org/10.2139/ssrn.3239047 . https://papers.ssrn.com/abstract=3239047 Accessed 2024-03-11

Johnston, L.G., Hakim, A.J., Dittrich, S., Burnett, J., Kim, E., White, R.G.: A Systematic Review of Published Respondent-Driven Sampling Surveys Collecting Behavioral and Biologic Data. AIDS and Behavior **20**(8), 1754–1776 (2016) https://doi.org/10.1007/s10461-016-1346-5 . Accessed 2024-03-11

Léon, L., Des Jarlais, D., Jauffret-Roustide, M., Le Strat, Y.: Update on respondent-driven sampling: Theory and practical considerations for studies of persons who inject drugs. Methodological Innovations **9**, 2059799116672878 (2016) https://doi.org/10.1177/2059799116672878 . Publisher: SAGE Publications Ltd. Accessed 2024-03-11

Mobarak, A.M., Sharif, I., Shrestha, M.: Returns to International Migration: Evidence from a Bangladesh-Malaysia Visa Lottery. American Economic Journal: Applied Economics **15**(4), 353–388 (2023) https://doi.org/10.1257/app.20220258 . Accessed 2024-04-07

Magnani, R., Sabin, K., Saidel, T., Heckathorn, D.: Review of sampling hard-to-reach and hidden populations for HIV surveillance. AIDS **19**, 67 (2005) https://doi.org/10.1097/01.aids.0000172879.20628.e1 . Accessed 2024-03-11

Naderifar, M., Goli, H., Ghaljaie, F.: Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. Strides in Development of Medical Education **14**(3) (2017) https://doi.org/10.5812/sdme.67670 . Publisher: Kerman University of Medical Sciences. Accessed 2024-03-11

Nosek, B.A., Hardwicke, T.E., Moshontz, H., Allard, A., Corker, K.S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M.K., Nuijten, M.B., Rohrer, J.M., Romero, F., Scheel, A.M., Scherer, L.D., Schönbrodt, F.D., Vazire, S.: Replicability, Robustness, and Reproducibility in Psychological Science. Annual Review of Psychology **73**(Volume 73, 2022), 719–748 (2022) https://doi.org/10.1146/annurev-psych-020821-114157 . Publisher: Annual Reviews. Accessed 2024-04-07

Nixon, S.: Of Rice and Men: Rice Consumption-Based Estimates of Undocumented Persons in Malaysia. International Migration Review **58**(1), 5–36 (2024) https://doi.org/10.1177/01979183221126466 . Publisher: SAGE Publications Inc. Accessed 2024-04-06

Home Affairs, M.: Current Policies on Recruitment of Foreign Workers

Putul, S.J., Mia, M.T.: Exploitation of Migrant Workers in Malaysia and Protection under Domestic Laws:. In: Proceedings of the International Law Conference, pp. 125–131. SCITEPRESS - Science and Technology Publications, Kuala Lumpur, Indonesia (2018). https://doi.org/10.5220/0010054801250131 . https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010054801250131 Accessed 2024-04-07

Pandi, A.R., Pereira, A.A., Yacob, A.: Assessment of Causes and Contributing Factors to Migrant Workers Becoming Undocumented in Malaysia. Technical report, International Organization of Migration, Geneva

Secretariat, A.: Research on Migrant Workers' Rights-Based Standard Employment Contract for Domestic Work in ASEAN. Technical report, ASEAN Secretariat, Jakarta

Slough, T., Tyson, S.A.: External Validity and Meta-Analysis. American Journal of Political Science **67**(2), 440–455 (2023) https://doi.org/10.1111/ajps.12742 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12742. Accessed 2024-03-11

Tabellini, M.: Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration. The Review of Economic Studies **87**(1), 454–486 (2020) https://doi.org/10.1093/restud/rdz027 . Accessed 2024-04-07

Theng, T.T., Muhamad Noor, N., Khalidi, J.R.: COVID-19: We Must Protect Foreign Workers. Technical report, Khazanah Research Institute, Kuala Lumpur (June 2020)

Yi, S., Simler, K., Loh, W.S., Tan Wei, K.: Who is Keeping Score? Estimating the Number of Foreign Workers in Malaysia. Technical Report 148289, World Bank Group, Washington DC (April 2020)