SIINC data analysis report

Ari Bronsoler

Ph.D. in Economics candidate

M.I.T

Executive summary.

This report has been performed by an individual field researcher to document and verify Social Impact Metrics generated by the SIINC collaboration between Roots of Impact and Clínicas del Azúcar (CDA) and provide additional insights on the impact that the SIINC is having on the patients and the clinics. The report finds 4 main conclusions:

1. The reports presented by CDA are accurate and their data processes reliable.
2. There is a causal effect of the SIINC on the Bottom of the Pyramid (BOP) proportion of patients in CDA. The effect is estimated to be 0.02, which represents an increase of 6% relative to the 0.34 baseline levels.
3. The SIINC seems to improve the clinics' ability to attract BOP patients without impairing its ability to attract non-BOP members.
4. Patients that continue treatment at CDA improve their health a lot. On average, they reduce their HbA1c by 2 points and the effect does not disappear even after 24 months of treatment.

This report is structured based on the 4 points above as reflected in the table of contents below.

Table of contents

1.- Reports audit.

This section discusses the process through which I verified the data that CDA reported on their reports. The researcher can confidently conclude that the reports delivered by CDA were honest and accurate. The section is organized as follows: Understanding the report, data sources, replication of results and conclusion.

1.1.- Understanding the report

To fully understand how the reports were made, the researcher visited Monterrey on the second week of October. The intention of this visit was to be able to observe first-hand how CDA manages its data and manipulates it toward creating their reports. The visit included several productive meetings with Ricardo Londono, CDA's CIO, to understand data management and a 2-hour meeting with the clinic's CEO Javier Lozano in which he explained the methodology through which they build their reports.

During the meeting with the CIO, he explained that in order to be able to identify the socioeconomic status of each arriving patient, they purchased a dataset made by Inteligeo, a data consulting firm, that classifies each suburb (*colonia*) of the country into the following categories:

Socioeconomic level classification

Category	Meaning
AB	High
C+	Middle-high
C	Middle
D+	Middle-low
D	Low
E	Marginalized
SD	No determination
N/A	Not available

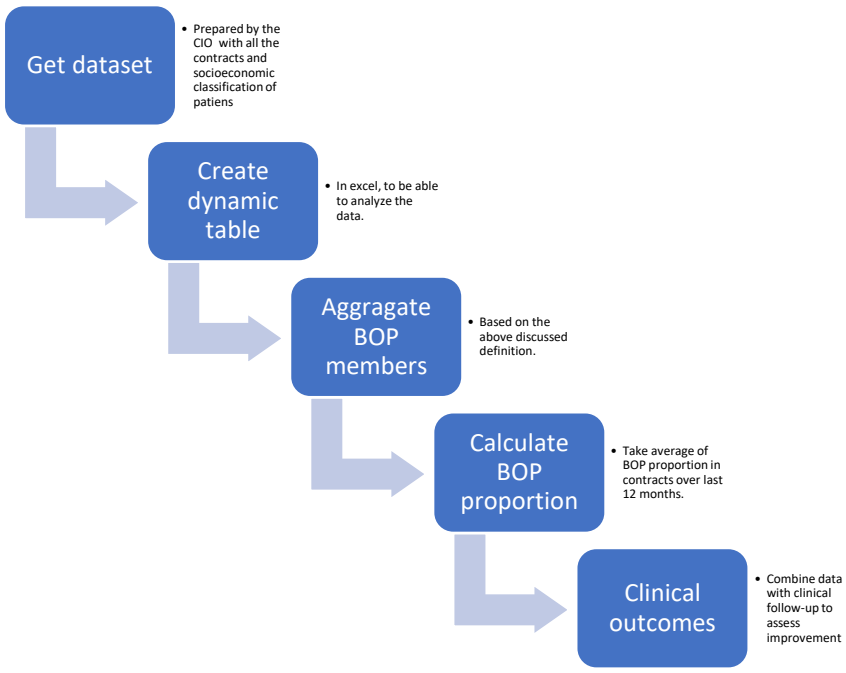
He further explained that the SD category should be considered as BOP because Inteligeo mentioned that their main clients are interested in middle-income populations and that when they are certain that a suburb will be of the lowest income they do not go and measure it. Therefore, all the SD suburbs are expected to be E classification, but it was not worth it for Inteligeo to confirm it. However, when there is no information available for a patient (N/A), Inteligeo has no information about where they live and therefore has no idea of that area's economic status. Therefore, the BOP population is defined as D or lower, including SD but excluding N/A. In my opinion, this is very sensible.

Now, in order to extract the BOP share of the population, they divide the BOP members over all members, including the N/A population. A potentially better way to extract the proportion of BOP members would have been to remove N/A patients from the total given they do not know their socioeconomic level and are potentially counting as not BOP patients some who are. However, this is not a big concern as there is only around 6% of patients without information so the bias is small. Further, such a bias would work against CDA finding an improvement given the BOP proportion would me artificially smaller. Lastly, and reassuringly, given the goals for the SIINC are based on baseline measures from CDA data, and the data process through which the KPIs are measured does not change, it is unlikely that another classification would have had different results.

Based on the above classification, the CEO creates and reports the BOP proportion at the SIINC clinic in Juarez and the evolution of such proportion at the other clinics as well as the progress in terms of HbA1c

that the BOP population experiences. One area of concern when evaluating the BOP proportion in the rest of the clinics is that potentially some BOP members stop going to a clinic because of the opening of Juarez. If Juarez is closer to their home, it is likely that they would rather go there instead of going to another. However, even with this limitation, the BOP proportion at the rest of the clinics seems to increase.

During the meeting with the CEO, he explained how he calculates the KPIs that CDA reports. He illustrated the process with a version of the data he had and how he manipulates it in order to construct the report. This proved to be a very helpful exercise as it allowed to observe how he builds the reports and to understand the rationale behind his process. A summary of the steps he takes below:



1. Open a dataset containing all contracts purchased at every clinic in Microsoft Excel. This dataset contains information on the socioeconomic status of each patient based on Inteligeo and allows the CEO to define BOP population.
2. Create a dynamic table allowing the tracking of the number of patients of each category every month by clinic.
3. Add the patients that are D or lower and S/D to define BOP and add the total number of patients by each clinic.
4. To report on BOP proportion, he calculates the proportion of BOP members over the last 12 months up to the 15th of the reporting month by adding up all patients from the BOP and dividing that over the total. He does this exercise for Juarez only and then repeats the process for every clinic except Juarez.
5. To analyze the progress that the BOP population has made in terms of HbA1c he uses another dataset that contains the HbA1c measurements for all the patients at the 6-month mark (defined as any follow-up measurement between 150-210 days). He averages over all the patients to acquire the measure reported.

Overall, this process is consistent. However, it is important to make sure that the data he gets to create the reports are created in a consistent manner. The next section reviews the raw data and talks about how the data was audited.

1.2.- Data sources and verifying information quality

In order to evaluate the data creation process, there were several meetings with the CIO, and he shared all the raw data he uses to create such reports in order to replicate the results. The data received came from direct extractions from their servers and had no human manipulation before being delivered. He shared three datasets. A summary below:

1) Inteligeo- a dataset containing a classification of socioeconomic levels for each suburb in Nuevo Leon, Mexico.
2) Contract data- a dataset containing every contract that was created in CDA that contains every patient's address. This dataset also contains the socioeconomic level that the CIO reports to the CEO to assess how accurate it is.
3) Laboratory data- a dataset containing every lab HbA1c measurement from every patient, along with the date of the measurement.

Based on the above datasets, the patients' addresses were matched to the Inteligeo dataset and the BOP proportions replicate. The datasets contain every contract since October 2015 and have 36,178 contracts in total. Out of the whole population, after combining the Inteligeo and contracts dataset by suburb, it is possible to observe the same socioeconomic level classification for 99.2% of the contracts. This gives a lot of confidence that the data on which the CEO bases his analysis is reliable and comes from combining the Inteligeo data with the patients' addresses. The next section presents a replication of the results CDA reported.

1.3.- Replicating results:

This section shows the results obtained from replicating the CEO's methodology using Stata based on the datasets obtained from the CIO. The results are extremely like the ones that CDA reported. On the attached "checking socioeconomic levels" and "checking patient evolution" do-files one can replicate all my processes in Stata. Results inspire confidence that the results presented by CDA are correct as the differences in indicators are very small. Further, the fact that such small differences exist reflects that the data provided was the immediate data from their system and not a previously manipulated version.

Metric 1: BOP proportion at Juarez- 12-month average on each period

Time	Baseline May-17		Period 1 Feb-18		Period 2 Aug-18		Period 3 Feb-19		Period 4 Aug-19	
	Report	Audit	Report	Audit	Report	Audit	Report	Audit	Report	Audit
Total BOP	N/A	N/A	289	279	549	549	556	553	N/A	523
Total Goal BOP	N/A	N/A	510	486	888	898	881	878	N/A	875
%	N/A	N/A	31%	31%	32%	32%	33%	33%	33%	33%
BOP %	N/A	N/A	56.7%	57.4%	61.8%	61.1%	63.1%	63.0%	N/A	59.8%
Diff			**-0.7%**		**0.7%**		**0.1%**		**N/A**	

Metric 2: BOP proportion at all the other clinics- 12-month average on each period

Time	Baseline May-17	Period 1 Feb-18	Period 2 Aug-18	Period 3 Feb-19	Period 4 Aug-19

	Report	Audit	Report	Audit	Report	Audit	Report	Audit	Report	Audit
Total BOP	1,938	2,976	2,826	2,906	2,833	2,841	2,857	2,861	N/A	2,785
Total	2,160	8,420	8,137	8,300	7,892	7,894	7,866	7,860	N/A	7,734
Goal BOP%	N/A	N/A	34%	34%	34%	34%	34%	34%	34%	34%
BOP %	N/A	N/A	34.7%	35.0%	35.9%	36.0%	36.3%	36.4%	N/A	36.0%
Diff			**-0.3%**		**-0.1%**		**-0.1%**		**N/A**	

Metric 3: HbA1c changes after 6 months

Time	Baseline May-17		Period 1 Feb-18		Period 2 Aug-18		Period 3 Feb-19		Period 4 Aug-19	
	Report	Audit	Report	Audit	Report	Audit	Report	Audit	Report	Audit
HbA1C change 6 months	N/A	2.2	2.2	2.3	2.5	2.4	2.4	2.3	N/A	2.8
Diff	**N/A**		**-0.10**		**0.08**		**0.09**		**N/A**	
HbA1c after 6 months	N/A	7.8	7.8	7.6	7.7	7.7	7.9	7.6	N/A	7.4
Diff	**N/A**		**0.19**		**-0.01**		**0.31**		**N/A**	

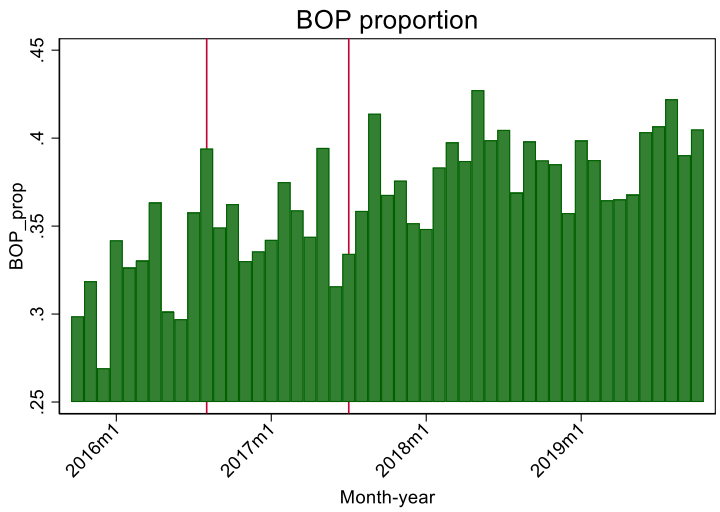
Overall, this shows that the reports presented by CDA are accurate. The next section focuses on estimating the causal effect on the BOP proportion at CDA.

2.- Causal estimation of the effect on BOP proportion.

Another thing of interest is to understand the effect that the SIINC intervention had on the overall BOP proportion. This is to some extent captured by the first 2 metrics, but in this section, a quantification of of the SIINC on these changes is attempted.

2.1 Descriptive statistics

This section describes how the data looks. There are 2 important dates for this evaluation. The first one is August 2016 when the SIINC term sheet was signed and the second one is July 2017 when the Juarez clinic opened. Below is the monthly BOP proportion evolution since October 2015.



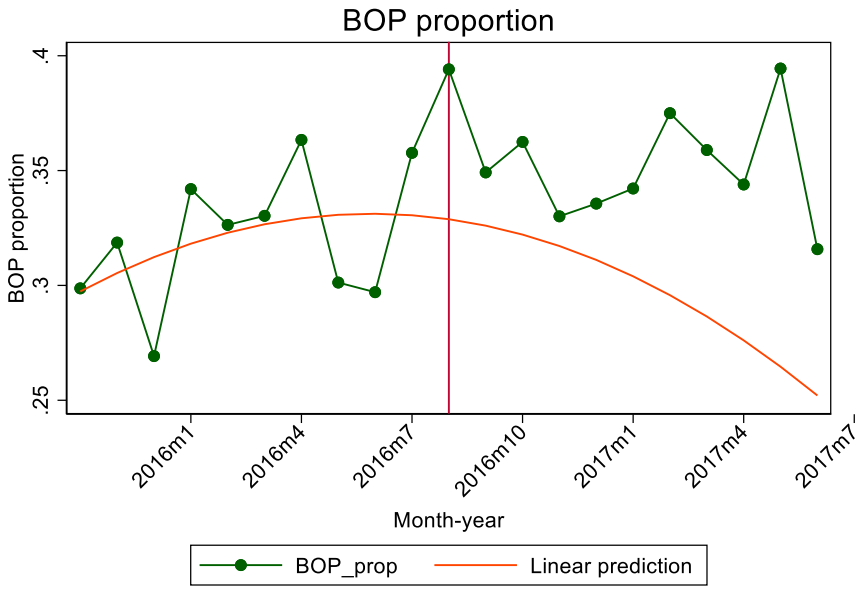
It is possible to see an increasing proportion with an increasing trend along with some seasonality. This makes it hard to attribute any changes to the SIINC intervention. In order to take this into account and be able to make a causal claim the empirical strategy below is utilized.

2.2.- Empirical strategy.

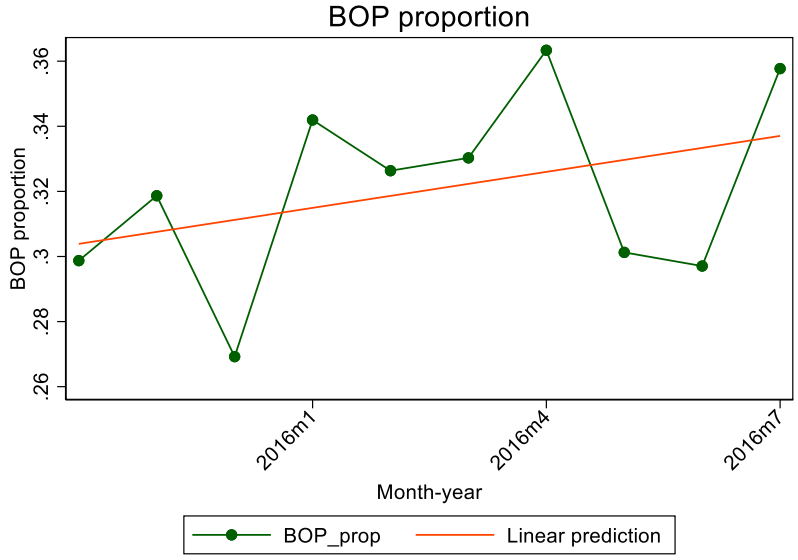
To assess the effect of the SIINC on the BOP proportion of the clinics, a prediction model was adjusted in the pre-SINC period with quarter fixed effects to account for seasonality and a quadratic monthly trend to allow the proportion evolution to have some flexibility. Afterward, predicted values from this model for the SINC period are used as a counterfactual. Lastly, it is possible to compare the actual rate to the counterfactual to assess the effect of the SIINC on the BOP proportion. This exercise is repeated taking as an initial SIINC date the signing of the sheet and the effect it had on the BOP share until Juarez opened and taking the initial date as the opening of the Juarez clinic and measuring its effect. The code for the analysis presented below can be found on the "Causal BOP proportion" do file attached. In the next two sections, deviations necessary the above strategy in order to get the most sensible counterfactual possible are explained.

2.2.1 Model for August 2016 start date.

For this model, there are only 10 months prior to the start of the program, so it is not possible include the month fixed effects because they would be colinear with the trend and would not allow us to see it. After trying to adjust with quarterly fixed effects, they had to be omitted as well because they were still too colinear to the monthly trend. Therefore, the prediction model is a linear trend because the quadratic model was over-adjusted, and the quadratic term was dominating the prediction (this happened probably because it was not possible to account for seasonal effects) as shown below when trying to take the prediction out of sample.

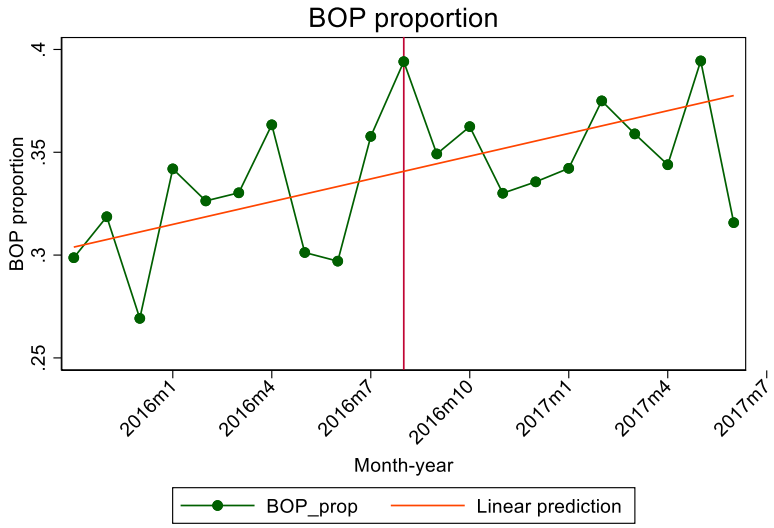


It is hard to think that the BOP proportion would have behaved like that, therefore, the results are based on the predictions of a linear model. The adjusted model's prediction for the period prior to the start looks like it is predicting well. The graph below presents how it looks versus the data where it is possible to see that it predicts well overall. Note that the adjustment is not much worse than the figure above for the pre-intervention period.



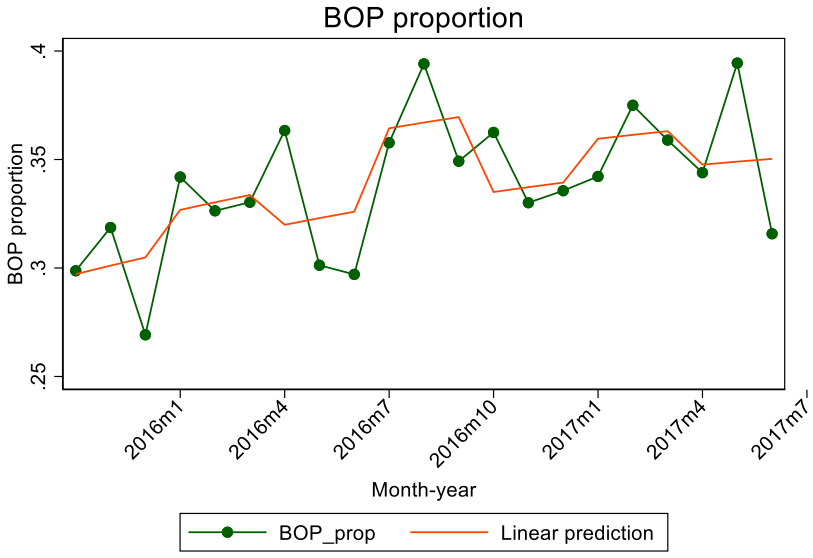
Now, in order to estimate the causal effect, it is necessary to assume that the evolution would have continued as our model suggests. In this case, the assumption is that that the BOP proportion evolves according to the linear trend presented above, which is the best option given the data limitations. After assuming this, if the model underestimates the BOP proportion (red line below green dots), then it is possible to conclude that the start of the SIINC program led to an increase in the BOP proportion.

The graph below reports how the model predicts for the months prior and after the intervention, it can be seen that there is no effect after Juarez's opening as the linear model continues to predict well. Therefore, the conclusion is that there was no effect on BOP proportion during this period.

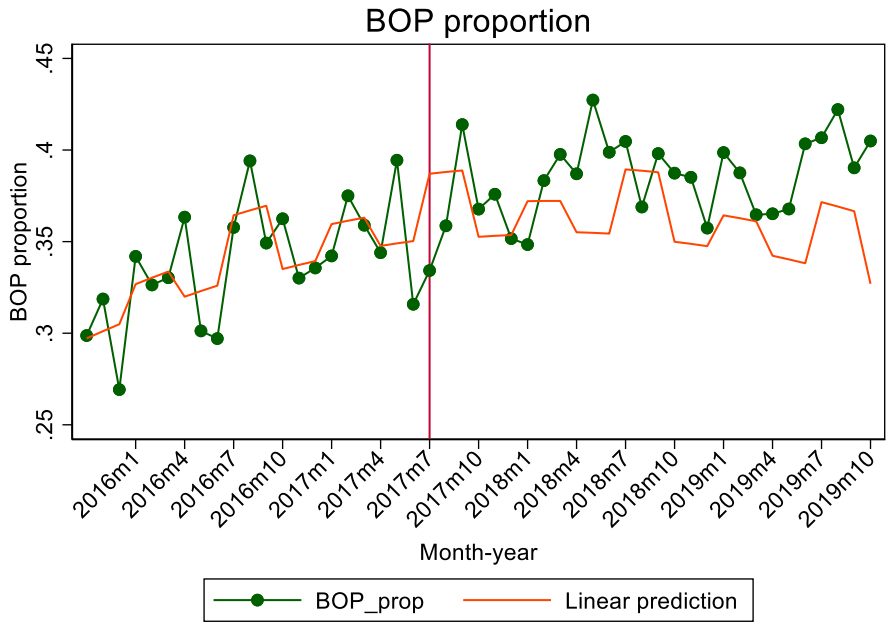


2.2.1 Model for July 2017 start date.

For this model, there are 21 months prior to the start of the program. However, since some months show up only once, it is still not possible to include the month fixed effects because they would perfectly predict that observation, which would be equivalent to removing it from the sample. However, it is possible to include quarter fixed effects. Therefore, this result is based on the predictions of a quadratic model with quarter fixed effects. The adjusted model's prediction for the period prior to the start looks like it is predicting well. The graph below presents how it looks versus the data, where it shows that it is predicting well.



Now, as above, in order to estimate a causal effect, it is necessary to assume that the evolution would have continued as the model suggests. If the model underestimates the BOP proportion, then it is necessary to conclude that the start of the SIINC program led to an increase in the BOP proportion. This is exactly what the graph below shows.



To quantify the effect of the program, the researcher takes the difference between reality and the prediction for before and after the intervention and run a linear regression of the differences on a constant and a dummy for after marking the start of the intervention. If there is an effect, the coefficient for after should be positive. This is exactly what the research finds, and the coefficient is significant at the 1% level. The estimated causal effect of the SIINC on the BOP proportion of patients at CDA is 0.02 (which represents an increase of around 6% relative to the pre-intervention mean of 0.34). The regression output is presented below.

```
. reg dif after
```

Source	SS	df	MS	Number of obs	=	49
				F(1, 47)	=	7.51
Model	.005466066	1	.005466066	Prob > F	=	0.0087
Residual	.034219555	47	.000728076	R-squared	=	0.1377
				Adj R-squared	=	0.1194
Total	.039685621	48	.000826784	Root MSE	=	.02698

dif	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
after	.0213426	.0077893	2.74	0.009	.0056726	.0370126
_cons	1.73e-18	.0058881	0.00	1.000	-.0118454	.0118454

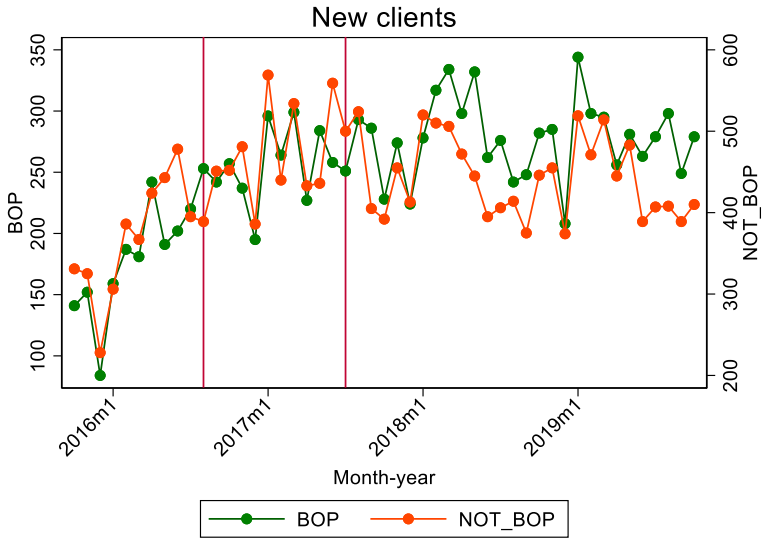
3.- Estimation of the effect on the total number of clients.

Another thing of interest is to understand the effect that the SIINC intervention had on the total number of clients at CDA as a reference of their potential for development and if it affected the clinics' strategy moving forward. On the one hand, the SIINC could enable the clinics to offer a better and more attractive service to BOP patients. However, on the other hand, it could limit the clinics' focus on their regular patients, leading to unintended consequences. This section shows that the SIINC increases the clinics' capabilities to treat BOP members without affecting their success on the rest of the population.

The analysis of this section can be replicated using the Clients evolution do file attached. Below is presented an overall time-series and a split by BOP members vs not-BOP members so that the reader can observe the effects. As above, the 2 red lines mark the possible dates of SIINC starting. Let us start with the overall graph. There seems to be a steady increase in the number of new clients as the SIINC start, which stabilizes after a while. This does not mean that the clinics have stopped growing, it just means that they are now growing at a steady rate.

The stabilization in growth might be a consequence of the economic climate of uncertainty that the country has been going through. Unfortunately, it is not possible to control for it in the analysis so instead the researcher will abstract from it. However, the effect of the SIINC on the clinics' work can be seen by splitting the time series into BOP vs non-BOP members and comparing their dynamic trends.

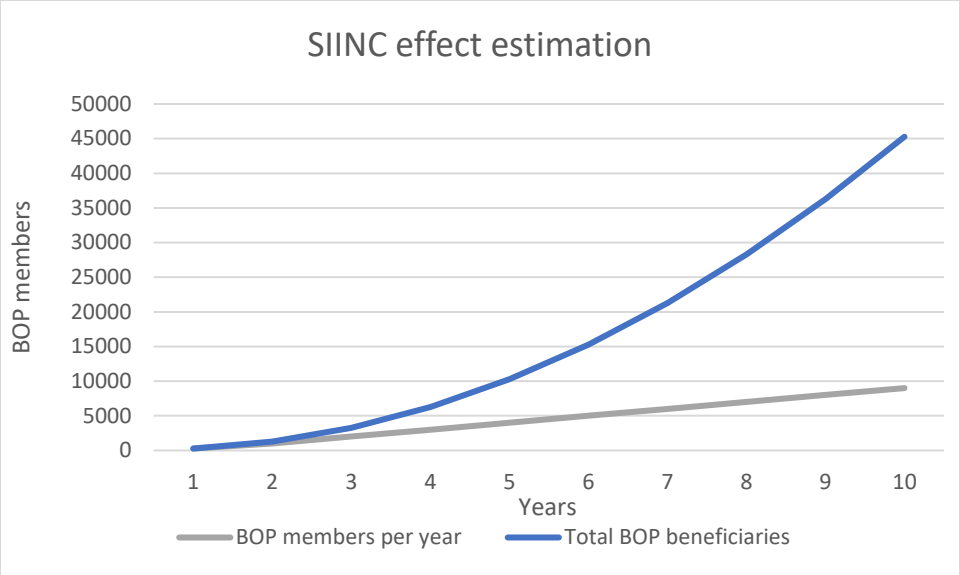
To see if the patterns differ, it is necessary to plot the evolution of new clients for both BOP and non-BOP patients. Before the SIINC, both had very similar evolution patterns, which is a nice feature that will allow for the capture of divergence clearly without having to de-trend the data. After the SIINC, there is an increase in the number of BOP members while there is no reduction in the new clients who are not BOP. This leads to the conclusion that the SIINC improved the clinics' ability to attract BOP patients without impairing their ability to treat not-BOP patients.



To further complement the analysis presented above, the researcher discussed with the CEO what his impression is of the SIINC intervention and how he thinks it has affected the clinics' strategy moving forward. He mentioned that the results so far have been remarkable and that he would have never thought that the BOP population would respond as well as they have to CDA. He explained that thanks to the SIINC, he is no longer afraid of opening clinics in more marginalized neighborhoods and that he is excited about incorporating this insight into the expansion strategy the clinics will experience over the next several years.

Now, what does that mean for the future as CDA expands. The CEO has made the claim that he intends CDA to have 200 clinics in the following 5 years. Therefore, assuming that the number of clients remains constant for each clinic (roughly 1,000 per clinic a year), and that the effect of the SIINC remains the same at 2%, it is possible to estimate that 4,000 more BOP patients will be treated by CDA thanks to the SIINC in year 5. This is a conservative estimate because the fact that the CEO intends to adapt the expansion strategy to the new insights on BOP population has not yet been incorporated.

The following graph shows the number of BOP patients that would benefit from the SIINC each year and the total number of BOP beneficiaries as time progresses when making the assumption that CDA will reach 200 clinics in 5 years and that they will continue the same growth rate for the next 5, while also assuming that the effect of the SIINC will remain at 2%.

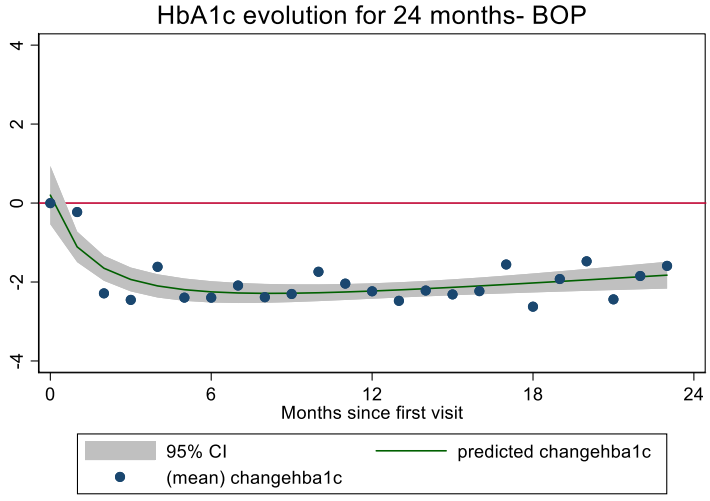


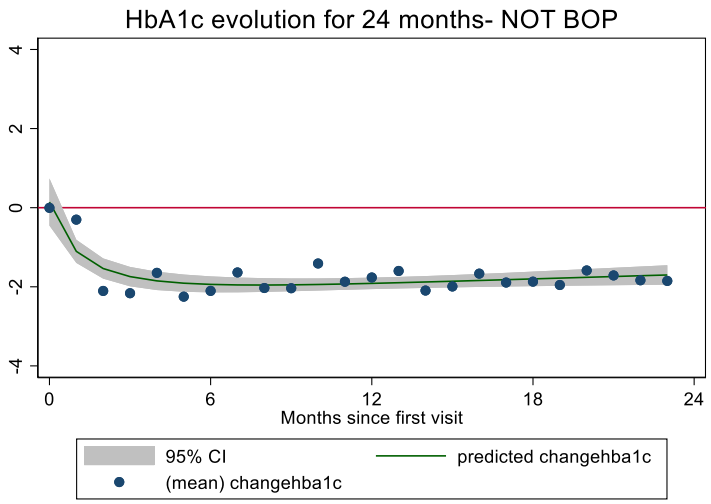
The next section presents descriptive statistics about patients' improvement in HbA1c.

4.- Descriptive statistics of how patients' health improved.

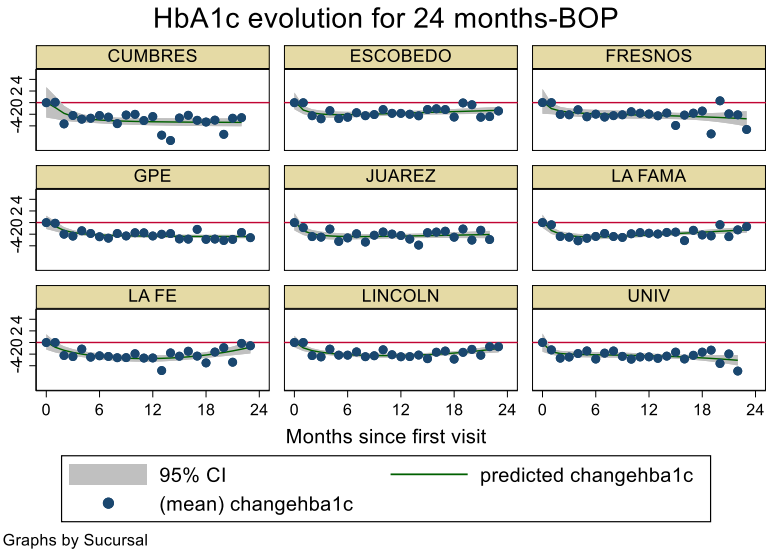
This section presents descriptive statistics of the effect that the clinics have on patients that continue treatment. The analysis for this section can be replicated with the Descriptive HbA1c do file attached. The focus is on the key outcome: HbA1c. Both outcomes for the BOP population are analyzed and then compared to the rest of the patients. Before getting into the analysis, it is important to highlight that the BOP patients start treatment at a very unhealthy point. On average, they have an HbA1c measure of 10, which is slightly worse than the rest of the population which has a baseline HbA1c average of 9.5.

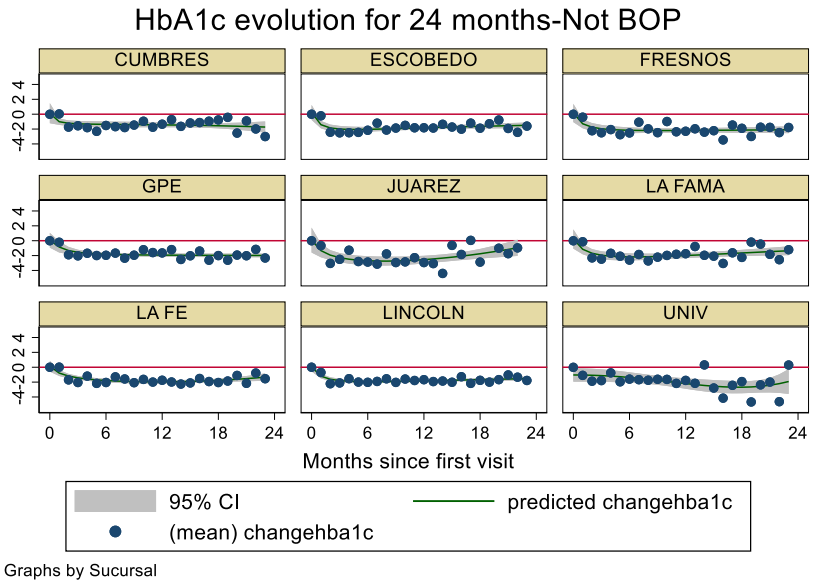
The graph below shows the progress made on average by the BOP patients and the rest. There is a marked improvement right after starting treatment and that this improvement does not disappear as time goes by, which is extremely hard to find in the literature and reflects CDA's great service. Further, there are no visible differences in progress between both groups.





Below the analysis is further split by clinic. It is possible to see that the effect is homogeneous across clinics and that Juarez (with a higher BOP concentration) does not show different patterns for either BOP or not BOP members.





Overall, these graphs highlight that patients that stay with CDA do better in terms of health.

Concluding remarks

The report finds 4 main conclusions:

1. The reports presented by CDA are accurate and their data processes reliable.
2. There is a causal effect of the SIINC on the Bottom of the Pyramid (BOP) proportion of patients in CDA. This effect is estimated to be 0.02, which represents an increase of 6% relative to the 0.34 baseline levels.
3. The SIINC seems to improve the clinics' ability to attract BOP patients without impairing its ability to attract non-BOP members.
4. Patients that continue treatment at CDA improve their health a lot. On average, they reduce their HbA1c by 2 points and the effect does not disappear even after 24 months of treatment.