Does national health insurance promote better indirect child outcomes: heterogeneous impact from estimation methods

Dennis Lim

Singapore Management University

July 2021

Abstract

This paper investigates whether national health insurance improves child health outcomes. More importantly, it compares between different estimation techniques in the presence of small sample sizes. To do this, we exploit the Demographic Health Survey data 2017-2018 for Pakistan to analyse the impact of the prime minster national health insurance scheme implemented in 2015. We apply propensity score matching (PSM) to account for differences in distribution between treated and control groups. Sensitivity analysis was conducted to evaluate robustness of our results. The results are mixed – success of the NHIS is dependent the health outcome measured as proxy. Outcomes that are significant are also robust to hidden bias, while outcomes that are insignificant are not. Differences within the class of PSM methodologies result in heterogeneous treatment estimation. In the most balanced matching method, treatment estimation is robust to alternative treatment estimation techniques outside of the PSM class, such as the inverse probability of treatment weighing estimator (IPTW). We also find that the more balanced the matching method, the closer it approximates the IPTW, suggesting application of IPTW in absence of large sample sizes to estimate average treatment effect of the treated (ATT). We also confirm that average treatment effect (ATE) performs poorly when approximated by PSM methods in small samples. Such evidence is relevant for both the understanding of health impacts and small sample approximation.

Keywords: Child Health, Health Insurance, Pakistan

1. Introduction

There has been varied accounts suggesting the significance of the effectiveness of national insurance. For instance, national health insurance has been shown to improve health outcomes in Egypt (Rashad et al 2019), India (Aggarwal 2010), Taiwan (Lee et al 2010) and Philippines (Quimbo et al 2011) while this effect is not apparent in China (Lei and Lin 2009) or Costa Rica (Dow and Schmeer 2003). This mixed result is reiterated by Levy and Meltzer (2001) who compared between various observational studies conducted, showing little evidence as to the causal relationship between national health insurance and health outcomes. The mixed result is largely due to the lack of randomization in observational data that inevitably leads to biasness in estimation (Quimbo et al 2011). At a closer look, Dow and Schmeer (2003) applied instrumental variables (IV) accounting for fixed effects but does not consider if the applied dataset was balanced. Smith and Todd (2005) suggested that the use of difference-in-difference estimator (DID) matching estimator performed best due to elimination of the presence of unobserved characteristics. Lei and Lin (2009) applied this method to their analysis of health insurance in China. However, their methodologies were questionable. For one, their classification of observations into treated and controls could not differentiate precisely or with certainty that the treated group were in fact treated. Second, the matching estimator did not ensure balancing of covariates, a vital requirement for the treated effects literature (Dehejia 2005). Upon consideration of the problems giving rise to insignificant health outcomes as described, there is a consensus that health insurance does improve health outcomes. Hadley (2003) considered the range of health research conducted over the past 25 years and concluded that despite the varying degrees of methodological flaws present in various researches, there exist substantial consistency in the result of health insurance across the board.

Past research in Pakistan's health development has shown that lower social groups generally lack financial resources to obtain private sector health services and hence avoid using any health services (Mumtaz et al 2013). National health insurance was aimed at reducing user fees to allow access to health services for the poorer households (Nyman 1999). While most of the Pakistan health research literature considers only the impacts of community campaigns and health insurance on direct health outcomes of children, such as immunization (Jamal et al 2020, Habib et al 2017) and child labour (Landmann and Frolich 2015), to the best of my knowledge, there has been no research done to evaluate indirect health outcomes of children, such as maternity care. The recent national health insurance implemented in Pakistan (2015) allows us to produce pioneering work in this area. The first aim of our paper is therefore to evaluate the impact of the newly minted national health insurance in the context of child's indirect health outcomes in Pakistan.

In the papers discussed above (paragraph 1), the number of treated observations range from 1555 to 2268, which is generally about 10% of the sample size at the minimum. This compares to 620 and 50,495 treated and untreated observations respectively in our dataset when we consider Pakistan's national health insurance (1.23% treated proportion compared to untreated), a reflection of treated population proportion at the point of data collection. It has been noted that in small samples, there may be insufficient power to produce meaningful inferences (Quigley 2003). This may explain why no research has yet been done on the evaluation of Pakistan's national health insurance – due to implementation being very recent leading to small treated samples. In such circumstances, it is recommended that a priori variables that could be strongly related to

outcome be included in the propensity score model (Brookhart et al 2006). To the best of my knowledge, only one paper has evaluated the efficiency of different estimation techniques in the presence of limited treated sample (Anais et al 2020) based on a simulation study - they found that the average treatment of the treated (ATT) was susceptible to variabilities from different propensity score matching methods and advised for sensitivity analysis to be conducted, although no technique to overcome this issue was developed. There is thus a deeper need to develop methods to overcome this limited treated-sample problem from an observational standpoint. Despite the shortcoming that there has been no theoretical research to overcome this issue, we can still apply real life observational data to the context of Anais et al (2020) to verify the hypothesis they provided – this will be the second aim of our paper: evaluating the robustness of treatment effect under several propensity score matching algorithm and comparing it against alternative estimators to determine its performance. This paper will hence contribute to the existing literature by providing empirical underpinnings for future work when researchers are faced with limited treated samples.

We first address the issue that treated group may differ from the untreated group in terms of distribution of unobserved variables, which may lead to biasness in estimation of treatment effects by making use of propensity score matching (Dehejia and Wahba 2002). This is done through estimating the probability of enrolment into national health insurance for any given household, accounting for a plethora of factors that would likely affect membership (Brookhart et al 2006). Second, we will assess the impact of the prime minister national health insurance programme (PMNHIP) on indirect health outcomes of children, such as number of antenatal visits as well as number of tetanus injection a mother takes before the birth of her child. We apply sensitivity analysis to evaluate the robustness of our results. This is applicable only in the context of average treatment of the treated (ATT). Third, we apply inverse probability treatment weighing (IPTW) as another estimator for additional robustness as well as obtaining the average treatment effect (ATE).

The results of the paper are first that the PMNHIP improves indirect health outcomes for children whose households are insured in Pakistan, depending on measured outcome. Even though the number of antenatal visits improves by 0.153 while the number of tetanus injection improves by 0.283 - which amounts to a 5.7% and 20.7% increase respectively compared to baseline controls - only the latter outcome (tetanus injection) is significant at the 10% significance level. This result is further cemented by a sensitivity analysis that shows that the number of tetanus injection is relatively insensitive to unmeasured confounding (Γ=1.25 at the 10% significance level). Second, based on the suggestions made by Anais et al (2020), we provide evidence contrary to their recommendation. Indeed, despite the insensitivity of tetanus injection, the results are heterogeneous within the class of propensity score matching (PSM) models in the presence of small sample. Comparing against PSM models that possess the balanced covariates however leads better results, highlighting the importance of comparing against only balanced sets (Harder et al 2010). Third, the lower the biasness, measured from Rubin's B value, the closer the treatment effect (for ATT) is to IPTW. [1] This suggests use of IPTW in the estimation of both ATT and ATE in the future when researchers face small sample issues.

[1] I would like to investigate this relationship more in the future, but due to paucity of time, this will be dealyed

The remainder of our paper will be structured as follows. Section 2 provides background information on Pakistan prime minister's national health insurance programme. Section 3 provides the description of the data. Section 4 covers the empirical strategy. Section 5 explains the result of our findings while section 6 discusses the robustness of our results. Section 7 concludes.

2. Healthcare in Pakistan and Prime Minister's National Health Insurance Programme (PMNHIP)

National health planning in Pakistan started as early as 1965, covering family planning programs, disease prevention programs such as malaria eradication programme, Expended program for immunization (EPI) and tuberculosis control programs under the National Institute of health (NIH) up till 2000 (Mashhadi et al., 2016). In 2001, decentralization of government authority aiming to increase healthcare delivery - in part from the United Nations Millennium Development Goals (UNMDG) signed in 2000 - led to the establishment of provincial level healthcare departments, up to the formation of Ministry of National Health Services, Regulation and Coordination in June 2011. However, out-of-pocket spending is estimated to be at PKR 315 billion (USD$1.9 billion) in period 2011-2012, with *Punjab* district having the highest share of spending (see fig 1 below).[2]

Figure 1 *- Gross out of pocket health expenditure in 2011-12 by region*[3]

Province/Area	Billion Rs.	% share
Punjab	171,355	54
Sindh	75,145	24
KPK & FATA	49,795	16
Balochistan	16,168	5
Islamabad	2,370	1
Total	314,833	100

On 31 December 2015, the Pakistani government launched the PMNHIP to provide basic healthcare to families making less than USD$2 a day identified under Benazir income support program (BISP) database. This would take effect over several phases starting with a targeted 23 districts on 1 January 2016, covering over 3 million families living below the poverty line, eventually progressing to cover more than 21 million (10% of total Pakistani) households by 2030.[4] Since the launch of the program, more than 6.7 million families have been enrolled.

The program is segregated into two treatment packages. The first is for general treatment that covers PKR 60,000 (USD$358)/year/household in Patient Services (All Medical and Surgical Procedures), emergency treatment requiring admission, maternity services (Normal delivery and caesarian section), maternity consultation, immunization, fractures and injuries, post hospitalization, local transportation and provision of transport to tertiary care hospitals. The second covers up to PKR 300,000 (USD$1790) /year/household

[2] Naturally, any national health insurance would initially target a larger proportion of the Punjab region, which we observe following the launch of PMNHIP

[3] Source: Pakistan national health accounts 2011-2012, Pakistan bureau of statistics. Available at http://www.pbs.gov.pk/sites/default/files/national_accounts/national%20health%20accounts/NHA_Report_2011-12.pdf

[4] https://www.pmhealthprogram.gov.pk/districts-covered-by-the-program/

for more immediate medical care, such as heart diseases, diabetes, hepatitis, HIV, organ failure, cancer, chronic infections and neurological procedures.[5]

The success of the program depends on the outcomes generated by its coverage, such as immunization and maternal consultation outcomes. Therefore, we require a dataset that measures a variety of outcomes covered by the PMNHIP, possesses an extensive array of factors that could potentially explain differences in these measured outcomes as well as measured sometime after the program's implementation. These criteria lead us to the next section.

3. Data and descriptive statistics

3.1 Data sources

The main data used in this paper is obtained from the world bank Demographic and Health Survey (DHS) microdata library for year 2017-2018, which is the fourth of its survey conducted for Pakistan. Even though the PMNHIP was introduced only recently and the full effects of its implementation can be more clearly seen after longer periods, due to the limitations of time, we will be contented with the 2017-2018 dataset for now.[6] This is sufficient for our needs mainly because it allows us to compare the effects of the PMNHIP before and after treatment (2015). Being an extremely comprehensive survey, it allows us to account for covariates such as households' education, number of children five and under, religion, number of household members and other indirect exogenous factors that could potentially affect health outcomes.

Figure 2[7] Map of Pakistan divided into major regions

The survey is also representative at the national level, containing eight main regions in Pakistan (Punjab, Sindh, Balochistan, Gilgit-Baltistan, Khyber Pakhtunkhwa, Islamabad capital territory, Azad Jammu and Kashmir, federally administered Tribal areas), shown in fig 2 above. At a sub-level, the data indicated whether the chosen individual lived in an urban or rural area. Furthermore, the data segregates individuals by whether it is a

[5] https://www.pmhealthprogram.gov.pk/about-us/

[6] The 2023-2024 survey dataset that will be available in the next few years can be used for better examination of policy effectiveness

[7] Image from https://en.wikipedia.org/wiki/Districts_of_Pakistan

Capital/large city, small city, town, or countryside. The data accounts for heterogenous population sizes across districts over a two-stage stratified sampling method. The first stage involves selecting sample points (clusters) consisting of Enumeration Blocks (EB). These EBs were then drawn with a probability proportional to their size, which is based on the number of households residing in that particular EB at the time of sampling. Upon choosing the EBs, the second stage involves sampling 28 households in every cluster. Each household in a fixed cluster had an equal probability of being chosen. This clustering is desirable as it allows us to match observable pre-treatment characteristics between the treatment and control group, to be discussed in the *methodology section*.

Also inclusive in the DHS is a wealth index for each household, which is a composite measure of household wealth based on ownership of certain asset such as types of water that is accessible to the individual, access to sanitation facilities, television and bicycles and materials used in the construction of housing.[8] This is convenient for our purposes as it allows a more comprehensive measure of wealth rather a specific measure of each characteristic.

3.2 Variables

3.2.1 Health insurance

The main dependent variable is a dummy variable equal to one if the *household has the national insurance* card (Sehat Sahulat program) and zero if the household does not have any insurance. Although there are other forms of health insurance, such as employer provided insurance, they do not possess the marginal effects we are looking for in evaluating healthcare provided to the poorer households.[9]

3.2.2 Anthropometrics for children's health

A drastic neonatal mortality decline of about 70% has been linked to an increase to "at least two" tetanus injection as well "at least three" antenatal-care visits (Singh et al 2019). Women who had tetanus injections before birth was also less likely to experience infant mortality at a significant level (Makate M & Makate (2017). As such, we consider the number of tetanus injection a mother receives before the birth of her child as a first proxy for child health's outcome.

Antenatal care also has a significant impact on pregnancy outcomes (Chari et al 2019). Furthermore, it was suggested that antenatal care improves the infant's quality of life in the long run (Almond and Currie 2011). The fetal programming literature has argued that antenatal care would in the later years affect the infant's health such as significant decreases in cardiovascular diseases (Kuh et al 2007), capital accumulation and productivity, skills and fatherhood (Heckman 2007). Following the work of established papers in birth-related impacts (Levine et al 2016, Okeke et al 2020), we consider the number of antenatal visits as the second proxy of child's health outcome.

3.3. Descriptive statistics

Observations with none of the above health outcomes were first dropped from the dataset; if at least one health outcome was present in the observation, that observation was kept. In addition, we removed Balochistan due to there being zero observations of those who

[8] Construction of index found here: https://dhsprogram.com/pubs/pdf/DHSG4/Recode7_DHS_10Sep2018_DHSG4.pdf

[9] Individuals with insurance covered by non-national insurance will not be considered – A total of 6,175 observations were dropped

were insured and from this region. Islamabad-capital territory (ICT) was also dropped from the dataset due to perfect collinearity when running the logit regression discussed in section 5.1 later. This results in 60 insured and 6,191 uninsured observations remaining in the dataset. We will consider these remaining observations to be the starting block of our analysis going forward.

3.3.1 Health outcomes

Table 1 summarizes the health outcomes of the unmatched dataset. Based on sample data's averages, the population of insured mothers experience 20% less antenatal visits compared to their uninsured counterparts at 5% significance level, indicting a highly unbalanced dataset pertaining to this outcome. However, the number of tetanus injections received before birth by an insured mother is slightly higher by around 3.6% compared to an uninsured mother, at an insignificant level.

3.3.2 Covariates

Table 2 summarizes information on the child, mother, father and household's characteristics. The gender of the child as well as proportion of household head being a male does not vary significantly between the insured and un-insured households. There are around 38% more non-educated mothers in the insured group than in the uninsured group at the 1% significance level, leaving the proportion of insured mothers having primary, secondary and higher education lower than the uninsured mothers at every level. We observe the same thing happening for fathers. The proportion of un-educated fathers who are insured is around 85% higher than the uninsured fathers with no education, leaving the proportion of insured fathers with primary, secondary and higher education to be lower than uninsured fathers at every level. As educated parents are likely to earn more, they generally do not qualify for PMNHIP, which requires a daily earning of less than US$2.

The proportion of insured families who stay in an urban area is about 12% lower than an insured family. There is a higher proportion of insured staying in Khyber Pakhtunkhwa (KPK) and Galgit-Baltistan (GB) than uninsured households; this leaves a lower proportion who stays in Punjab, Sindh, Azad Kashmir (AJK) and federally administered tribal areas (FATA). These differences in proportion were all statistically significant at the 1% level. However, there does not seem to be a significant relationship between proportion of being insured and wealth quantile; Punjab has on average the highest wealth percentile (3.48), followed by KPK (2.96) and AJK (2.76). In general, insured families' have 0.8 lesser household members, has less children who are below the age of 5, and are generally poorer. The wealth quintile is given from one to five, with one being the poorest and five the richest.

Table 1
Descriptive statistics – health outcome

	Insured			Not insured			Difference	
	N	Mean	s.d.	N	Mean	s.d.	Diff	(t-stat)
No. of antenatal visit	60	3.2	2.313	6,190	4.00	3.06	-0.800**	(-2.66)
No. of Tetanus	60	1.65	1.400	6,191	1.59	1.24	-0.058	(0.321)

*p<0.1, **p<0.05, ***p<0.01; sampling weights used to compute averages; s.d.: standard deviation; N = number of observations

Table 2
Descriptive statistics - covariate

	Insured		Not-insured		Difference	
	Mean	s.d.	Mean	s.d.	Diff	(t-stat)
Child's characteristics						
Child is male	0.583	0.497	0.514	0.500	0.069	(1.07)
Age of child in single years	1.783	1.519	1.451	1.317	0.332*	(1.69)
Mother's education						
Education: None	0.733	0.446	0.532	0.498	0.200***	(3.46)
Education: primary	0.15	0.36	0.208	0.406	-0.0584	(-1.25)
Education: Secondary	0.0167	0.129	0.108	0.310	-0.0916***	(-5.34)
Education: Higher	0.100	0.302	0.150	0.358	-0.051	(-1.285)
Father's characteristics						
Male head of household	0.917	0.279	0.887	0.316	0.0294	(0.812)
Education: None	0.467	0.503	0.252	0.434	0.214***	(3.29)
Education: primary	0.117	0.323	0.146	0.353	-0.0295	(-0.702)
Education: Secondary	0.267	0.445	0.373	0.483	-0.107**	(-1.85)
Education: Higher	0.15	0.360	0.228	0.419	-0.078	(-1.66)
Household's characteristics						
Urban	0.417	0.497	0.421	0.494	-0.051	(-0.079)
Region: Punjab	0.0833	0.279	0.254	0.435	-0.171***	(-1.69)
Region: Sindh	0.033	0.181	0.222	0.416	-0.189***	(7.88)
Region: KPK	0.617	0.490	0.204	0.403	0.413***	(6.50)
Region: GB	0.200	0.403	0.0875	0.283	0.112***	(2.15)
Region: AJK	0.033	0.181	0.128	0.337	-0.096***	(-4.06)
Region: FATA	0.033	0.181	0.103	0.303	-0.069***	(-2.92)
Number of household members	9.92	5.94	9.10	4.70	0.812	(1.06)
Number of children < 5 year	2.1	1.36	2.23	1.46	-0.132	(-0.744)
Wealth quintile	2.17	1.21	2.72	1.39	-0.553***	(-3.52)
N	60		6,191			

*p<0.1, **p<0.05, ***p<0.01; s.d.: standard deviation; N: number of observations; data only includes children that are currently alive

4. Empirical methodology

In any observational study (such as this), the lack of randomization generally leads to a systematic difference between the average treatment effect (ATE) and the average treatment of the treated (ATT).[10] As insurance status may be correlated to parents' education as well as determinants of health status, comparing insured households to un-insured households directly may lead to biased ATE. The use of propensity score matching (Rosenbaum and Rubin, 1983a) was proposed to overcome this issue. However, the *strongly ignorable treatment assignment* property must be satisfied in our setting to obtain the ATE estimate, which are: (a) the potential outcome of the individual should be independent from the treatment assignment given the observed covariates and (b) the probability of being treated given a vector of covariates must be positive and less than one. [11,12] The first condition is known as the "no unmeasured confounders" condition. We consider the Rosenbaum bounds sensitivity analysis (Rosenbaum 2002) in order to evaluate the extent of this condition's significance in altering our results below.

The propensity score is defined as $\mathbf{e(X_i) = Prob(Z_i = 1|X_i = x)}$, where X_i is a vector of covariates for individual i and $Z_i = 1$ if individual was treated. The balancing score $b(X_i)$ is defined as a function that satisfies $X_i \perp Z_i \mid b(X_i) = b$. Theorem 1 of Rosenbaum (and Rubin 1983) states that the propensity score is also a balancing score. In addition, conditional

[10] i.e. $E\{Y_i(1) \mid Z_i = 1\} \neq E\{Y_i(1)\}$, where $Y_i(0)$ denotes the outcome when individual i is assigned to the control group

[11] More precisely, $\{(Y_i(1), Y_i(0)) \perp Z_i\} \mid X_i = x$

[12] Mathematically, $0 < Prob(Z_i = 1|X_i = x) < 1$

on the true value of the propensity score, the distribution of baseline covariates will be independent of treatment assignment[13]. Together with theorem 4 of Rosenbaum (and Rubin 1983) - which states that for any dataset that possess the *strongly ignorable treatment assignment* property, given a balancing score, the expected difference between the treated and untreated subjects will be unbiased - we can obtain the treatment effects of health-outcomes between those with and without insurance. Therefore, in practice, we first solve for the propensity score. This propensity score is most generally obtained by applying a logistic regression, using a set of covariates (i.e. the propensity score being the probability of having health insurance in our case) as regressors.[14] Second, we require appropriate methods to determine if our propensity score model has been correctly specified, since misspecification implies that theorem 4 may not apply. Hence, our propensity score model should be checked for balancing. Given any fixed value for the true propensity score, the distribution of baseline covariate should be similar between the treated and untreated subjects. **Appendix 8.1** describes some matching estimators that will be applied in section 5. These estimators reflect different methods for achieving a balanced sample; balancing tests will be used to determine the matching algorithm with the smallest bias (Austin 2009)[15]. This is further explained in **Appendix 8.3**.

5. Econometric analysis

5.1 Determinant of household insurance membership

Previous research investigating determinants of healthcare membership in Pakistan suggests that parents' education as well as whether a household lives in a rural or urban district are significant factors (Toor et al 2005, Asif and Akbar 2020). Additionally, other papers have found that wealth, number of children and age of children to affects insurance membership (Amo 2014, Salari et al 2019). We hence estimate the propensity score using the following logistic regression:

$$I_i = \beta_0 + \beta_1 X_{Ci} + \beta_2 X_{Mi} + \beta_3 X_{Fi} + \beta_4 X_{Hi} + u_i$$

I_i is an indicator of whether household i is insured by the Prime minister National health insurance programme (PMNHIP); $X_{Ci}, X_{Mi}, X_{Fi}, X_{Hi}$ are respectively the covariates of household i in **Table 2** above, corresponding to child's characteristic, mother's education, father's characteristic and household characteristics. Child characteristics include gender and age (in single years), father's characteristics include education level and whether head of household is male; household characteristics include region, area (urban or rural), number of household members, number of children less than five years of age and wealth quintile.

Table 3 below provides the result of the logistic regression. The probability of having national insurance increases generally with a poorer educational background for both the father and the mother. Households with male child as well as older children has a higher probability of being insured, although at an insignificant level. In terms of region, the probability of insurance is significantly higher in Khyber Pakhtunkhwa (KPK) and Gilgit-Baltistan (GB). In addition, a lower number of children less than 5 years as well as a

[13] This is shown in appendix 8.2

[14] There are other methods used to estimate the propensity score, such as bagging (Lee et al 2010), tree-based networks (Setoguchi et al 2008) among many others

[15] You can read it for free at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3472075/

lower wealth leads to higher probability of being insured, at a 10% significant level. The probability of insurance also increases in the number of household members significantly.

Table 3
Probability of having health insurance

Variables	Coefficient	(s.e.)
Child's characteristics		
Male	0.276	(0.275)
Age in single years	0.109	(0.103)
Mother's education		
Primary	0.0273	(0.394)
Secondary	-1.481	(1.009)
Higher	0.149	(0.492)
Father's characteristics		
Male head of household	0.0508	(0.479)
Primary	-0.575	(0.429)
Secondary	-0.699*	(0.352)
Higher	-0.600	(0.461)
Household's characteristics		
Urban	-0.135	(0.293)
Region		
Punjab	0.531	(0.860)
Sindh	-0.707	(0.977)
KPK	2.544***	(0.730)
GB	2.075**	(0.755)
AJK	0.245	(1.009)
Number of household members	0.0722*	(0.0345)
Number of children < 5 year	-0.212*	(0.0904)
Wealth quintile	-0.319*	(0.139)
Constant	-5.347***	(0.776)
N	6,251	

*p<0.10, **p<0.05, ***p<0.01; Baseline categories are mother's education: none, father's education: none, region: FATA; s.e.: White"s robust standard error; N: number of observations. Dependent variable = 1 if household is covered under PM national health insurance (or = 0 if not insured at all)

5.2 Propensity score and matching algorithm

The propensity score model in the described in **table 3** satisfies the balancing property. In other words, households that have the same probability of having the prime minister's national health insurance programme will have similar distribution of covariates that are independent of the status of health insurance. We choose the matching method that minimizes overall bias among the group of matching algorithms which also satisfies an adequately balanced sample criteria according to Rubin (2001). The balancing tests shows that the K-nearest neighbour matching algorithm produced the most balanced sample and reduces overall bias by as much as 86.7% compared to the unmatched sample (**table 7**). The value of K = 5 is chosen as it provides the local minimum for Rubin's B value. The K-nearest neighbour matching procedure is explained in **appendix 8.1**. This matching algorithm leads to an adequately balanced sample (Rubin 2001) – given the propensity score, the distribution of covariates between the treated and untreated subjects are similar. Under this *K-nearest neighbour* algorithm, the support area of propensity score is large, with treatment group and control group possessing an estimated range of [0.0003674; 0.13126] and [0.0003635; 0.12735] respectively, compared to the interval of [0.0000382,0.12735] for the untreated group before matching. This common support (after matching) restricts comparisons only among observations that were matched, resulting in 60 treated observations and 248 untreated observations, with each untreated observation replaced if being drawn in previous match.

5.3 Impact of PMNHIP on health outcome

The health outcomes were based on matching on k-nearest neighbour algorithm, after which average treatment effect of the treated (ATT) was calculated by the method described under **appendix 8.1.** We can expect that if PMNHIP is successful, there should

be significantly positive ATT for both health outcomes under a balanced dataset, which we constructed in the previous section using propensity score matching (PSM).

Table 4 describes the impact of health insurance on health outcomes for children. Insured pregnant women have roughly 0.15 more antenatal visits compared to a pregnant woman in an uninsured household while insured pregnant women receives 0.28 more tetanus injections. The results suggest that having health insurance significantly (one-tailed) increases the number of tetanus injection by about 21%. In order to allow our causal inference to be valid, we apply the Abadie-Imbens (2002) standard error.

Our average treatment of the treated (ATT) estimates will only be valid under the assumption that there were no unmeasured confounders in the estimation of the probability model. In the presence of unobserved covariates, our inference may not valid due to biasness from PSM (Becker & Caliendo 2007). Hence, we conduct a Rosenbaum bounds sensitivity analysis for hidden biases. **Table 5** describes the range of Γ values that would render our analysis inappropriate; we demonstrate this by giving an example: if we assume Γ = 1.05, that will mean that for households with the same vector of covariates, the probability of being insured in the treated group is 1.05 times higher than the control group. The values under each Γ is the highest level of significance before treatment is insignificant (p-value). We can see that even in the presence of no hidden bias(Γ=1), the ATT for number of antenatal visits is already insignificant at the 10% level. For the number of tetanus injection, the ATT described in **table 4** is valid up to Γ=1.15 at the 5% significance level and Γ=1.25 at the 10% significance level, implying that the effect of health insurance in number of tetanus injection even in the presence of some confounding is valid.

ATE for both health outcomes are insignificant; we also observe a negative value for the number of antenatal visits, which can be explained by incomplete matching between the controls to the treated samples. This will be addressed using other estimators in the next section. The negative values obtained for the estimation of ATE for the number of antenatal visits using propensity score matching (PSM) suggests persistence of small sample biasness of the treated as well as an unbalanced sample set supported by **table 7**. To see this, based on the observed dataset, the probability of being insured (i.e. ($Prob(Z_i = 1)$) is calculated to be at 0.96%, while currently, about 21% of households in Pakistan are already insured by the program.[16] In addition, the unbalanced set has led to the sample mean of untreated outcome to be much lower than treated outcome. This combined effect led to the ATE estimation using PSM into the negative region, invalidating our ATE estimation using PSM.

Overall, our results suggest some improvement in health outcomes from insurance, even though the extent of its significance may be mixed. To see this, we compare the estimation of ATT using different matching methods described in **table 9 (Appendix 8.4)**. The number of antenatal visits generally does not differ much, while we observe that the number of tetanus injection remains significantly positive across matching algorithms to varying degrees. This suggests that our results can be somewhat sensitive to different matching algorithms. Several other authors have found this to be true as well, especially in the presence of disproportionate distribution of treatment population in sample sizes (Kurth et al 2006, Schafer & Kang 2008), although applied in other contexts from medicine to

[16] There are 6,750,306 families enrolled in the program now, out of 32.21million families in Pakistan - https://www.pmhealthprogram.gov.pk/district-enrollment-counts/

psychology. Other researchers suggest presenting treatment effect estimates from propensity scoring methods only if they possess the balanced covariate property (Harder et al 2010). Our results demonstrate that various propensity score matching methods, although balanced, can still lead to differences in estimated ATT in the presence of small sample size. This seems to suggest evidence for Kurth et al (2006)'s research more than Harder et al (2010) in the context of propensity score matching models (PSM).

Anais et al (2020) suggested the use of thorough sensitivity analysis in drawing conclusions using propensity score matching in the context of small sample sizes. Despite the number of tetanus injection being relatively insensitive to unmeasured confounding (results still hold at Γ=1.25), there is still some variability in ATT estimation among different PSM estimators. This directly contradicts the hypothesis that sensitivity analysis would be of help when comparing between PSMs. More care would have to be taken in the context of small treated sample for future researchers. Depending only on sensitivity analysis is insufficient indication of estimation's robustness to bias.

Table 4
Impact of health insurance on health outcomes from PSM

	N	Mean control	ATT	% effect	ATE
Number of antenatal visits	308	3.005	0.153	5.7	-1.223
			(0.330)		(0.240)
No. of Tetanus before birth	308	1.367	0.283*	20.7	0.600
			(0.194)		(0.328)

*p<0.1, **p<0.05, ***p<0.01; propensity score matching with K (=5) nearest neighbours algorithm; N: number of observations in common support area; ATT: average treatment of the treated as described in appendix 8.1; ATE: average treatment effect; parenthesis is Abadie-Imbens standard error for 5 nearest neighbours; % effect: |ATT|/| mean control|

Table 5
Sensitivity analysis

	Γ						
	1	1.05	1.1	1.15	1.2	1.25	1.3
Number of antenatal visits	0.1131	0.1472	0.18529	0.2268	0.2709	0.3168	0.3637
No. of Tetanus before birth	0.0169	0.0250	0.0355	0.0487	0.0645	0.0831	0.1044

Γ: hidden bias from Rosenbaum sensitivity analysis with 5 nearest- neighbour matching; p-values at each interval for each health outcome is provided

6. Alternative estimation techniques and robustness check

To address the disproportionate weighing from the PSM method that led to negative ATE results, we consider the inverse probability of treatment weighing estimator (IPTW). This will serve our purposes here two-fold. First, it allows for estimation of both ATT and ATE, unlike score matching methods, which generally allow only for the estimation of ATT (Harder et al 2010). This can serve as a robustness check regarding our estimation of ATT. We would expect that the ATT under different methods of calculation to be almost identical under a sufficiently large sample size. Since ATT does not consider the probability of being insured (unlike ATE), we can expect the IPTW to provide consistent estimates comparable to the PSM method.

Second, the IPTW overcomes the disproportionate weighing scheme for the ATE under the PSM method. Researchers sometimes choose to exclude certain observations that lie out of the propensity score of the treated observations (Heckman et al 1997), also known as the "common support region". In our data, the PSM limited the range of propensity score to [0.0003635; 0.12735] from [0.0000382,0.12735]. Even though the ATT estimation is still valid, it may render causal inference for ATE estimation invalid when only sub-

samples are used, leaving only ATE from IPTW valid in smaller samples (Austin & Stuart 2017).

The average treatment effect (ATE) and average treatment effect for the treated (ATT) in the presence of no unmeasured confounders using the IPTW estimator is given as:

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e(X_i)}$$

$$\mathrm{ATT_{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Z_i}Y_i - \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\frac{\mathrm{e(X_i)(1-Z_i)Y_i}}{1-\mathrm{e(X_i)}}$$

Where n is the number of observations in dataset, $e(X_i)$ is the propensity score discussed in section 4, estimated through a logistic regression on covariates[17].

We applied Wooldridge's analytical standard errors (WASE) on the ATE, while the standard bootstrapping procedure was used for the standard errors of ATT to account for variance attributed to estimation of score matching procedure (Heckman et al 1997, Sianesi 2004, Wooldridge 2010, Bagnoli 2019), due to the disproportionately low weighing on the probability of being insured (i.e. $(Prob(Z_i = 1))$ leading to an extremely high inverse weight in the calculation of standard errors for ATT based on WASE, which has led to invalid standard errors in the absence of large sample to reflect the population probability of being insured[18]. **Table 6** provide the results for ATE and ATT from the IPTW estimator. The results suggest that the number of antenatal visits would increase by about 0.7 if the PMNHIP was extended to the entire population, while the number of tetanus injection would increase by roughly 2.3 at a significant level. Among the group that was treated, there was a rough 0.12 increase in antenatal visits at an insignificant level, while the number of tetanus injections has increased by around 0.3 at a significant level.

A comparison between the results in **table 6** and **table 4** above suggests that our results are robust to some extent. We apply the unpaired t-test for equivalent ATT values between PSM and IPTW, which can be justified when the difference in observations (308 for PSM and 6250 for IPTW) is large enough such that we can assume both samples are independent. We first compare between the standard error for ATT for the two health outcomes between PSM and IPTW. Results show that at 10% significance level, there is insufficient evidence to reject the null hypothesis of no difference in standard deviation for both health outcome. We hence apply the equal variance t-statistics to compare between the ATT values from PSM and IPTW. Results show no significant difference in ATT values for both health outcome as well.

As the IPTW method allows all eligible observations to be used - unlike matching where certain incompatible observations are dropped to allow for a balanced sample - it avoids biasness potentially arising from incomplete matching, which can occur when some untreated observations are excluded from the result of different matching methods (Rosembaum & Rubin 1985). Our results therefore suggest that we can make some causal inferences from the ATT estimates provided in **table 4**, given that in the presence of biasness, we should notice a significant difference between PSM and IPTW in the estimation of ATT. In addition, we notice that the more balanced a matching method is (PSM), the closer the values of health outcome are to IPTW, suggesting that IPTW is a good proxy for ATT in the presence of small sample sizes. Our work suggests that more

[17] See appendix 8.5. for proof of unbiasedness of estimator. Alternatively, refer to pages 67-69 of Micro-Econometrics for Policy, Program and Treatment Effects by Myoung-jae Lee

[18] Section 3.1 of stata journal (2014), vol 14, number 3, pg 541-561 describes the calculation of WASE

theoretical work would have to be done to explain this phenomenon. At the minimum, we know at least that the most balanced propensity score matching estimator is robust to our IPTW estimation suggesting some reliability in our estimation results.

Table 6
Health outcome from IPTW

	N	ATT	ATE	var ratio test p-value (ATT)	Equal var t-test (ATT)
Number of antenatal visits	6,250	0.123	0.698**	0.3003	-0.926
		(0.282)	(0.345)		
No. of Tetanus before birth	6,251	0.256*	2.32***	0.1849	-1.08
		(0.184)	(0.189)		

*p <0.1, **p<0.05, ***p<0.01; propensity score generated from logit-regression model (as in 5.1); Standard error (in parenthesis) for ATE are calculated based on Wooldridge (2010, p. 922-924) for standard M-estimators, which provides asymptotically consistent standard errors[19]; Standard error for ATT based on a bootstrap of 500 replications; var ratio test p-value (ATT): variance ratio test using sdtesti command between ATT from PSM and IPTW; equal var t-test (ATT): 2 sample t-test for unpaired data

7. Conclusion

Despite mixed result in the impact of national health insurance in improving health outcomes, we show evidence that success is dependent on the specific health outcome evaluated. Implicit in previous papers dealing in health outcome evaluation is the existence of large number of observations within the treated sample. This paper contributes to the literature in being the first to evaluate the national insurance scheme in Pakistan as well as shed light on the issues related to robustness in small sample. We find that treatment outcome is highly sensitive to different PSM methods. We suggest using the most balanced matching method for analysis (lowest Rubin's B value). We note that sensitivity analysis is useful only outside the class of PSM estimators in that robustness will persist in results between the most balanced PSM estimation and alternative treatment estimation methods outside the class of PSM estimators. We also notice a general relationship between PSM class of models as well as alternative models, in that the more balanced a PSM method, the closer its treated estimation is to other estimation techniques, in our case, the IPTW estimator. Further research should investigate the existence of this relationship and how it pertains to treatment estimation in small sample. We also find that PSM method performs poorly when the estimand is the average treatment effect (ATE). Alternative estimation techniques should be used if the interest is in ATE. For one, IPTW works despite shortcomings of limited sample sizes.

[19] Visit https://fmwww.bc.edu/RePEc/bocode/t/treatrew.pdf for in-depth explanation

8. Appendix

8.1 *Matching algorithms for solving ATT*[20]

There are a select number of methods used for matching. We briefly discuss the different matching algorithms used in this paper followed by provision of average treatment of the treated estimator afterwards

1:1 matching

(i) *Nearest-neighbour matching* with and without replacement matches the untreated (or control) observation closest to the ith individual in terms of propensity score. If there are more than one closest matching of untreated observation to the treated observation, one of these untreated observations will be chosen at random. The only difference between with and without replacement is that after each matching between treated and control, the control observation that was used in matching will be replaced. $w_{ij} = \frac{1}{n_i}$ in the ATT context below.

(ii) *Caliper matching* (or radius matching) involves setting a fixed value for the radius of each treated observation, and only control observations in the fixed radius around that treated observation in terms of propensity score will be matched to it (Rosenbaum & Rubin 1985). $w_{ij} = \frac{1}{n_i}$ in the context below

(iii) *Kernel matching* involves simply setting weight between each treated and control observation inversely proportional to the distance between their propensity scores[21].

1:k matching

1:k neighbourhood matching with replacement, where k is fixed, involves taking each treated observation to match with k closest control observations that are closest to it in terms of propensity score. $w_{ij} = \frac{1}{n_i}$ in the context below

Finally, the estimator is given as:

$$ATT_{Propensity\ score\ matching} = \frac{1}{n_T} \sum_{i \in \{Z_i = 1\}}^{n_T} [y_{1i} - \sum_{j \in \{Z_j = 0\}}^{n_i} w_{ij} \,.\, y_{0j}]$$

Where $\sum_{j \in \{Z_j = 0\}}^{n_i} w_{ij} = 1$ for any $i \in \{Z_i = 1\}$; n_T is the number of treated observations; n_i is the number of control observations matched to the ith treated observation; y_{1i} is outcome of treatment for individual i, y_{0j} is outcome for untreated jth individual matched to the ith treated individual; w_{ij} is the weight assigned to jth untreated individual as a counterfactual to the ith treated individual. These weights are specific to the ith individual and may change respectively with different individuals.

[20] Stata software package for matching found here: http://repec.org/bocode/p/psmatch2.html

[21] Read more on pg 364 of https://journals.sagepub.com/doi/pdf/10.1177/1536867X0200200403

8.2 *Proof that outcome of treatment assignment is independent from distribution of baseline covariate conditional on propensity score:*

Treatment Z_i is a binary value, to show that $Z_i \perp X_i \,|e(X_i) = e$, we first show that $E[Zi = 1 \,|Xi = x, e(X_i) = e] = E[Z_i = 1 \,|e(X_i) = e]$ [i.e. $Prob(Z_i = 1|X_i = x) = Prob(Z_i = 1|e(X_i) = e)$]

First, note that $E[Zi = 1 \,|Xi = x\,, e(X_i) = e] = E[Zi = 1|Xi = x] = e$. Furthermore, $E[Zi = 1|e(X_i) = e] = E[E(Zi = 1 \,|e(Xi) = e, Xi = x)|e(Xi) = e] = E[E(Zi = 1|Xi = x)|e(Xi) = e] = E[prob(Zi = 1 \,|Xi = x)|e(Xi) = e] = E\{e(Xi) = e \,|e(Xi) = e] = e$. The first result follows.

Next, we note that $Prob(Z_i = 1 \,|X_i = x) = Prob(Z_i = 1 \,|X_i = x, e(X_i) = e) = \frac{Prob(Z_i = 1, X_i = x,\ e(X_i) = e)}{prob(X_i = x,\ e(X_i) = e)}$. Hence, $Prob(Z_i = 1 \,|X_i = x).prob(X_i = x, e(X_i) = e) = Prob(Z_i = 1, X_i = x, e(X_i) = e)$; we are now in a position to show that $Prob(Z_i = 1, X_i = x \,|e(X_i) = e) = Prob(Z_i = 1 \,|e(X_i) = e).Prob(X_i = x|e(X_i) = e)$: from the left-hand-side, $Prob(Z_i = 1, X_i = x \,|e(X_i) = e) = \frac{Prob(Z_i = 1, X_i = x,\ e(X_i) = e)}{prob(e(X_i) = e)} = Prob(Z_i = 1|X_i = x).\frac{Prob(X_i = x, e(X_i) = e)}{Prob(e(X_i) = e)} = Prob(Z_i = 1|e(X_i) = e).Prob(X_i = x|e(X_i) = e)$, which completes the proof

8.3 Standardized differences for balancing tests (or % bias)[22]

For *continuous variable*, the standardized difference is defined as

$$d_{continuous} = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

Where $\bar{x}_{treatment}$ *and* $\bar{x}_{control}$ denotes the sample mean of each covariate between the treated and untreated subjects, respectively. $s^2_{treatment}$ *and* $s^2_{control}$ denote the sample variance for each corresponding covariate between treated and untreated subjects, respectively.

For *dichotomous* variables, the standardized treatment is defined as

$$d_{dichotomous} = \frac{\hat{\rho}_{treatment} - \hat{\rho}_{control}}{\sqrt{\frac{\hat{\rho}_{treatment}(1 - \hat{\rho}_{treatment}) + \hat{\rho}_{control}(1 - \hat{\rho}_{control})}{2}}}$$

Where $\hat{\rho}_{treatment}$ *and* $\hat{\rho}_{control}$ represents mean of the sample for the dichotomous variable for treated and control variable respectively. The standardized difference compares the difference in means in units of pooled standard deviation. Matching of propensity score greatly reduces this standardized difference. Rubin's B, which is the absolute standardized difference of the means of the linear index of the propensity score between treated and untreated subjects. Rubin (2001) suggests this standardized difference to be less than 25 for samples to be considered adequately balanced.

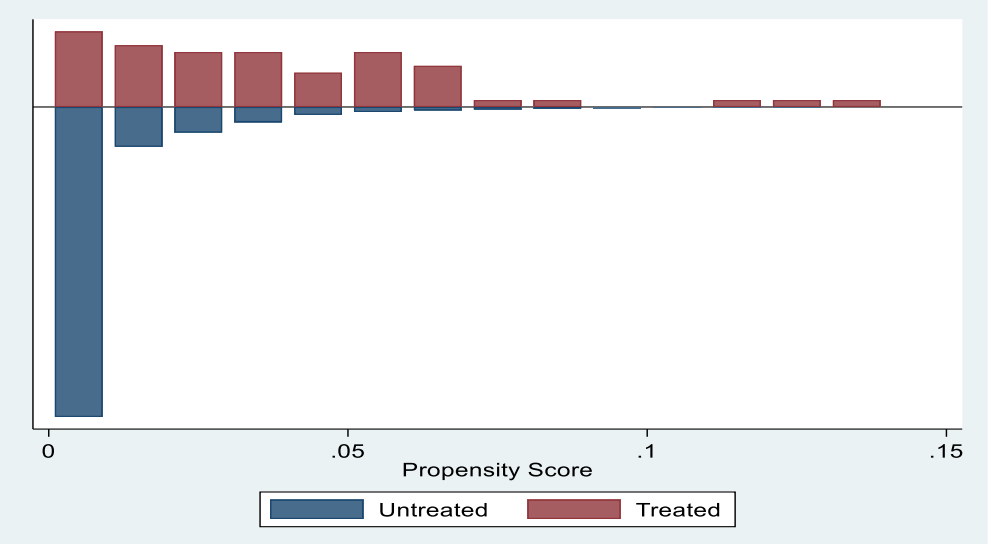
[22] Based on Austin (2009)

Table 7
Balancing tests

	% bias					
	Before matching	NN w/ repl	NN w/o repl	radius	Kernel	KNN w/ repl
Child's characteristics						
Male	13.8	-13.4	-13.4	2.7	8.9	0
Age in years	23.4*	0	-7.0	4.3	14.6	6.8
Mother's education						
Primary	-15.2	8.7	8.7	6.1	-10.3	10.4
Secondary	-38.5**	-7.0	-7.0	-3.2	-26.6	-4.2
Higher	-15.3	15.1	15.1	1.4	-10.4	2.0
Father's characteristics						
Male head of household	9.9	-11.2	-11.2	-0.5	6.7	4.5
Primary	-8.7	-9.8	-9.8	3.8	-5.6	-1.0
Secondary	-23.1*	3.6	0	3.1	-15.6	2.1
Higher	-19.9	4.3	4.3	1.3	-13.6	-2.6
Household's characteristics						
Urban	-1.0	26.9	20.2	-7.0	-1.1	9.4
Region						
Punjab	-46.7***	0	0	1.5	-31.5*	-2.7
Sindh	-58.9***	0	0	-2.7	-40.7**	3.1
KPK	92***	-11.1	-7.4	-1.3	62.2***	-2.2
GB	32.3***	14.4	9.6	1.0	22.5	0
AJK	-35.7**	0	0	0.1	-24.3	1.2
Number of household members	15.2	17.7	13.4	-14.3	10.6	4.6
Number of children < 5 year	-9.3	23.6	22.4	-15.2	-5.3	6.6
Wealth quintile	-42.5***	-7.7	-7.7	0.2	-28.7	-5.1
Average bias	27.8	9.7	8.7	3.9	18.8	3.8
% fall in bias from matching		65.1	68.7	86.0	32.4	86.7
Rubin's B	139.6	58.8	53.5	24.9	89.0	24.6

*p<0.1, **p<0.05, ***p<0.01; NN w/repl: nearest neighbour with replacement: NN w/o repl: nearest neighbour without replacement; radius caliper = 0.001; KNN w/repl: K-nearest neighbor with replacement; k = 5

Table 8
Histogram of propensity scores – before matching



8.4. Robustness checks

Table 9

Robustness checks – ATT from alternative matching algorithms

	Baseline: KNN (5) w/ repl	NN w/ repl	NN w/o repl	Radius [0.001]	Kernel
Number of antenatal visits	0.153	0.137	0.267	0.178	-0.503
	(0.330)	(0.542)	(0.456)	(0.317)	(0.284)
No. of Tetanus before birth	0.283*	0.383*	0.400**	0.341*	0.127
	(0.194)	(0.258)	(0.258)	(0.196)	(0.186)

*p<0.1, **p<0.05, ***p<0.01; NN w/repl: nearest neighbour with replacement: NN w/o repl: nearest neighbour without replacement; radius caliper = 0.001; KNN(5) w/repl: 5-nearest neighbor with replacement

8.5 Proof that inverse probability of weighing estimator (IPTW) is unbiased under no unmeasured confounders

We want to show that $\boldsymbol{ATE_{IPW} = E[Y(1)] - E[Y(0)]}$. Recall $ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-e(X_i)}$

It suffices to show that $E\left[\frac{Y_i Z_i}{e(X_i)}\right] = E[Y(1)]$ and $E\left[\frac{Y_i Z_i}{1-e(X_i)}\right] = E[Y(0)]$.

This can be seen from:

$$E[\frac{Y_i Z_i}{e(X_i)}] = E[E\left[\frac{Y_i Z_i}{e(X_i)}\middle| X_i = x\right]] = E[\frac{E[Y_i Z_i | X_i = x]}{e(X_i)}] = E\left[\frac{E[Y(1) Z_i | X_i = x]}{e(X_i)}\right]$$
$$= E\left[\frac{E[Y(1)|X_i = x]E[Z_i|X_i = x]}{e(X_i)}\right] = E\big[E[Y(1)|X_i = x]\big] = E[Y(1)]$$

Similarly,

$$E[\frac{Y_i(1-Z_i)}{1-e(X_i)}] = E[E\left[\frac{Y_i(1-Z_i)}{1-e(X_i)}\middle| X_i = x\right]] = E[\frac{E[Y_i(1-Z_i)|X_i = x]}{1-e(X_i)}] = E\left[\frac{E[Y(0)(1-Z_i)|X_i = x]}{1-e(X_i)}\right]$$
$$= E\left[\frac{E[Y(0)|X_i = x]E[1-Z_i|X_i = x]}{1-e(X_i)}\right] = E\big[E[Y(0)|X_i = x]\big] = E[Y(0)]$$

For the second part, we show that $\boldsymbol{ATT_{IPW} = E[Y(1) - Y(0)|Z_i = 1]}$

Recall that $ATT_{IPW} = \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{e(X_i)}{1-e(X_i)}(1-Z_i)Y_i$. It suffices to show that $E[Z_i Y_i] = E[Y(1)|Z_i = 1]$ and $E\left[\frac{e(X_i)}{1-e(X_i)}(1-Z_i)Y_i\right] = E[Y(0)|Z_i = 1]$.

First, $E[Z_i Y_i] = E[Z_i Y_i | Z_i = 1] + E[Z_i Y_i | Z_i = 0] = E[Y(1)|Z_i = 1]$

Second,

$$E_{x,z}\left[\frac{e(X_i)Y_i(1-Z_i)}{1-e(X_i)}\right] = E_x\left[E_{z|x}\left[\frac{e(X_i)Y_i(1-Z_i)}{1-e(X_i)}\middle| X_i = x\right]\right] = E_x\left[\frac{e(X_i)E_{z|x}[Y_i(1-Z_i)|X_i = x]}{1-e(X_i)}\right]$$
$$= E_x\left[\frac{e(X_i)E_{y,z|x}[Y(0)(1-Z_i)|X_i = x]}{1-e(X_i)}\right]$$
$$= E_x\left[\frac{e(X_i)E_{y|x}\big[Y(0)|X_i = x\big]E_{z|x}[1-Z_i\big|X_i = x]}{1-e(X_i)}\right] = E_x\left[e(X_i)E_{y|x}[Y(0)|X_i = x]\right]$$
$$= E_x\left[prob(Z_i = 1|X_i = x)E_{y|x}[Y(0)|X_i = x]\right]$$
$$= E_x\left[E_{z|x}[Z_i|X_i = x]E_{y|x}[Y(0)|X_i = x]\right] = E_x\left[E_{y,z|x}[Y(0)Z_i|X_i = x]\right]$$
$$= E_x\left[E_{y|x}[Y(0)Z_i|X_i = x, Z_i = 1] + E_{y|x}[Y(0)Z_i|X_i = x, Z_i = 0]\right]$$
$$= E_x\left[E_{y|x}[Y(0)|Z_i = 1, X_i = x]\right] = E_y[Y(0)|Z_i = 1]$$

$$Q.E.D$$

References

A. V. Chari, Peter Glick, Edward Okeke, Sinduja V. Srinivasan - Workfare and infant health: Evidence from India's public works program; Journal of Development economics, Vol. 138, May 2019

ABADIE, A., AND G. IMBENS (2002): "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," Technical Working Paper T0283, NBER

Aggarwal A. (2010). Impact evaluation of India's 'Yeshasvini' community-based health insurance programme. Health economics, 19 Suppl, 5–35.

Almond D, Currie J. Killing Me Softly: The Fetal Origins Hypothesis. J Econ Perspect. 2011;25(3):153-172.

Amo, T. (2014). The National Health Insurance Scheme (NHIS) in the Dormaa Municipality, Ghana: Why some residents remain uninsured? Global Journal of Health Science, 6(3), 82.

Andrillon Ananis, Pirracchio Romain, Chevret Sylvie (2020). Performance of propensity score matching to estimate causal effects in small samples. Statistical Methods in Medical Research, 29(3), 644–658.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine, 28(25), 3083–3107.

Austin, P. C., & Stuart, E. A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Statistical methods in medical research, 26(4), 1654–1670.

David Levine, Rachel Polimeni, Ian Ramage, Insuring health or insuring wealth? An experimental evaluation of health insurance in rural Cambodia, Journal of Development economics, Vol 119, March 2016

Diana Kuh, Yoav Ben-Shlomo: A life course approach to chronic disease epidemiology; Oxford University press, 2004

DiPrete, Thomas & Gangl, Markus. (2004). Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments. Sociological Methodology. 34. 271-310.

Edward N. Okeke, Isa S. Abubakar - Healthcare at the beginning of life and child survival: Evidence from a cash transfer experiment in Nigeria; Journal of Development economics, Vol 143, March 2020

Hadley J. (2003). Sicker and poorer--the consequences of being uninsured: a review of the research on the relationship between health insurance, medical care use, health, work, and income. Medical care research and review: MCRR, 60(2 Suppl), 3S–112S.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, *15*(3), 234–249.

Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. The Review of Economic Studies, 64(4), 605–654

Hussain, S., Hussain, R., Hafeez, A., & Khan, A. (2018). Prime minister's national health programme (PMNHP): A cost comparison analysis. Pakistan Journal of Public Health, 8(1), 37-42.

Jamal, D., Zaidi, S.A., Husain, S., Orr, D.W., Riaz, A., Farrukhi, A.A., & Najmi, R. (2020). Low vaccination in rural Sindh, Pakistan: A case of refusal, ignorance or access? *Vaccine*.

James J. Heckman, Hidehiko Ichimura, Petra E. Todd, Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme, The Review of Economic Studies, Volume 64, Issue 4, October 1997, Pages 605–654

James J. Heckman: The economics, technology, and neuroscience of human capability formation; Proceedings of the National Academy of Sciences Aug 2007, 104 (33) 13250-13255

Jeffrey M Wooldridge, 2010. "Econometric Analysis of Cross Section and Panel Data," MIT Press Books, The MIT Press, edition 2, volume 1, number 0262232588.

Jeffrey Smith and Petra Todd, (2005), Does matching overcome LaLonde's critique of nonexperimental estimators?, Journal of Econometrics, 125, (1-2), 305-353

John Nyman, (1999), The value of health insurance: the access motive, Journal of Health Economics, 18, (2), 141-152

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. American journal of epidemiology, 163(3), 262–270.

Landmann, Andreas & Frölich, Markus, 2015. "Can health-insurance help prevent child labor? An impact evaluation from Pakistan," Journal of Health Economics, Elsevier, vol. 39(C), pages 51-59.

Lee B.K., Lessler J., Stuart E.A. Improving propensity score weighting using machine learning. Statistics in Medicine. 2010; 29:337–346.

Lee, Y. C., Huang, Y. T., Tsai, Y. W., Huang, S. M., Kuo, K. N., McKee, M., & Nolte, E. (2010). The impact of universal National Health Insurance on population health: the experience of Taiwan. *BMC health services research*, *10*, 225.

Lei, X., & Lin, W. (2009). The New Cooperative Medical Scheme in rural China: does more coverage mean more service and better health? Health economics, 18 Suppl 2, S25–S46.

Levy H, Meltzer D. What do we really know about whether health insurance affects health? In: McLaughlin CG, editor. Health Policy and the Uninsured. Urban Press; Washington DC: 2004.

Lisa Bagnoli: Does health insurance improve health for all? Heterogeneous effects on children in Ghana; World Development, Vol. 124, December 2019

M. Alan Brookhart, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, Til Stürmer, Variable Selection for Propensity Score Models, *American Journal of Epidemiology*, Volume 163, Issue 12, 15 June 2006, Pages 1149–1156

Makate, M., & Makate, C. (2017). The Impact of Prenatal Care Quality on Neonatal, Infant and Child Mortality in Zimbabwe: Evidence from the Demographic and Health Surveys. Health Policy and Planning, 32(3), 395–404.

Muhammad Asif, Atta & Akbar, Muhammad. (2020). Inequalities in child health care in Pakistan: measurement and decomposition analysis of maternal educational impact. Public Health. 183. 94-101. 10.1016/j.puhe.2020.03.029.

Muhammad habib, Sajid Soofi, S. Cousens, Saeed Anwar, Najib U.H (2017) - Community engagement and integrated health and polio immunisation campaigns in conflict-affected areas of Pakistan: a cluster randomised controlled trial; VOLUME 5, ISSUE 6, E593-E603, JUNE 01, 2017

Mumtaz, Zubia & Levay, Adrienne & Bhatti, Afshan & Salway, Sarah. (2013). Signalling, status and inequities in maternal healthcare use in Punjab, Pakistan. Social science & medicine (1982). 94. 10.1016/j.socscimed.2013.06.013.

Nishtar S, Bhutta ZA, Jafar TH, et al. Health reform in Pakistan: a call to action. Lancet. 2013;381(9885):2291-2297

Paul R. Rosenbaum & Donald B. Rubin (1985) Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score, The American Statistician, 39:1, 33-38

Peter C. Austin - An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies; multivariate behavioral research; 2011 May; 46(3)

Quigley D.D. (2010) - Using multivariate matched sampling that incorporates the propensity score to establish a comparison group. CSE Technical Report No. 596. Los Angeles, California: University of California at Los Angeles, Center for the Study of Evaluation

Quimbo, S. A., Peabody, J. W., Shimkhada, R., Florentino, J., & Solon, O. (2011). Evidence of a causal link between health outcomes, insurance coverage, and a policy to expand access: experimental data from children in the Philippines. *Health economics*, *20*(5), 620–630.

Rajeev Dehejia and Sadek Wahba, (2002), Propensity Score-Matching Methods For Nonexperimental Causal Studies, The Review of Economics and Statistics, 84, (1), 151-161

Rajeev Dehejia, (2005), Practical propensity score matching: a reply to Smith and Todd, Journal of Econometrics, 125, (1-2), 355-364

Rashad, A. S., Sharaf, M. F., & Mansour, E. I. (2019). Does Public Health Insurance Increase Maternal Health Care Utilization in Egypt? Journal of International Development, 31(6), 516–520

Rosenbaum P.R. (2002) Sensitivity to Hidden Bias. In: Observational Studies. Springer Series in Statistics. Springer, New York, NY

Rosenbaum P.R., Rubin D.B. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society, Series B. 1983b;45:212–218.

Rosenbaum P.R., Rubin D.B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician. 1985; 39:33–38

Rosenbaum P.R., Rubin D.B. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983a; 70:41–55

Rubin, D.B. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. Health Services & Outcomes Research Methodology 2, 169–188 (2001).

Salari, P., Akweongo, P., Aikins, M., & Tediosi, F. (2019). Determinants of Health Insurance Enrolment in Ghana: Evidence from Three National Household Surveys. Health Policy and Planning, 34(8), 582–594.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. Psychological Methods, 13(4), 279–313.

Sehhatie F, Najjarzadeh M, Zamanzadeh V, Seyyedrasooli A. The effect of midwifery continuing care on childbirth outcomes. *Iran J Nurs Midwifery Res*. 2014;19(3):233-237.

Setoguchi S., Schneeweiss S., Brookhart M.A., Glynn R.J., Cook E.F. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. Pharmacoepidemiology and Drug Safety. 2008; 17:546–555

Sianesi, Barbara. (2004). An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s. The Review of Economics and Statistics. 86. 133-155.

Singh, A., Kumar, K., & Singh, A. (2019). What Explains the Decline in Neonatal Mortality in India in the Last Three Decades? Evidence from Three Rounds of NFHS Surveys. Studies in Family Planning, 50(4), 337–355.

Syed Fawad Mashhadi, Saima Hamid, Rukhsana Roshan, Aisha Fawad - HEALTHCARE IN PAKISTAN–A SYSTEMS PERSPECTIVE, Pak Armed Forces Med J 2016; 66(1):136-42

Toor, Imran Ashraf, and Muhammad Sabihuddin Butt - determinants of health care expenditure in Pakistan; Pakistan Economic and Social Review, vol. 43, no. 1, 2005, pp. 133–150

Volha Lazuka: The long-term health benefits of receiving treatment from qualified midwives at birth, Journal of Development economics, vol. 133, July 2018, https://doi.org/10.1016/j.jdeveco.2018.03.007

William Dow and Kammi K. Schmeer, (2003), Health insurance and child mortality in Costa Rica, Social Science & Medicine, 57, (6), 975-986

Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D (2010) - Use of stabilized inverse propensity score as weight to directly estimate relative risk and its confidence intervals. Value in Health. 2010; 2(13):273–7.