Does insurance affect health care utilisation if the health system is polarized? Evidence from a South African natural experiment

Abstract

We formulate and test a hypothesis that insurance only affects health care utilisation in countries where the health care system is sufficiently polarized that there exists a dramatic gap in the quality and characteristics of the care that can be accessed with and without insurance cover. We investigate this hypothesis in the context of South Africa, where apartheid era policies left the country with a highly unequal society and a polarized health care system. In order to estimate the causal effect of health insurance, we exploit the exogenous variation in health insurance coverage induced by the implementation of the Government Employees Medical Scheme. Two sets of data are used to test firstly the effect of the initial implementation of this policy in 2007 and secondly the effect of the continued roll-out between 2008 and 2012. Our identification strategy uses aspects of difference-in-difference and instrumental variable estimators to identify the causal effect of health insurance on health seeking behaviour. The results indicate that health insurance has a large effect on utilisation amongst South Africans, and that this effect is mainly due to the increased usage of high quality medical services. This result is found to be highly robust to the choice of dataset, sample period, specification and estimator. Estimates of the marginal treatment effect confirm that there is a very high pent-up demand for quality health care in South Africa, and indicate that health insurance is higher amongst individuals who expect to gain more from this. Further extensions in access to health insurance should therefore eventually produce somewhat weaker health utilisation responses.

JEL CODES: I11, I18, C31

1. Introduction

The demand for health services is notoriously difficult to disentangle from the demand for health insurance. In any voluntary health insurance environment, individuals tend to self-select into health insurance based on self-assessed risk and expected utilisation, leading to an overrepresentation of individuals who are cautious, sick or wealthy amongst the insured. This may induce an artificial positive correlation between health visits and insurance, which confounds the estimation of the causal effect of insurance using observational data.

A number of experimental studies have attempted to disentangle the relationship by allocating insurance or eligibility for insurance exogenously, often via a lottery, thus enabling causal estimations of how insurance affects health service utilisation and provider decisions (Card, Dobkin and Maestas, 2008; Finkelstein et al., 2012; Levine et al., 2014; Thornton et al., 2010). These studies examine health care choices when insurance facilitates access to better health services and/or lowers the cost of existing services. In developing countries there is reliable and robust evidence that insurance influences provider choice and out-of-pocket expenditure, but with elusive public health benefits because there is frequently no impact on health care utilisation. America is the exception. Experiments with insurance show a significant and notable impact on utilisation. We propose that this may be attributable to the polarized American health system and the consequent large gap between the type of health care that is accessible with and without health insurance.

Against this backdrop, we consider the launch of the Government Employees Medical Scheme (GEMS) in South African in 2006 as a natural experiment in expanding insurance coverage in a polarized health system. Under GEMS all government workers became eligible for health insurance subsidies and low earning employees received a full subsidy for the most basic benefit package, Sapphire. Under this scheme, no co-payments were required when using network providers. GEMS has had a dramatic impact on South Africa's health insurance landscape. Between 2006 and 2012 while there was little growth in the rest of the health insurance market, GEMS provided health care cover to 370 000 previously uninsured households (Moloabi, 2013). Since its launch GEMS has generated a steady stream of enrolments and by 2013 GEMS was still receiving 7000 monthly applications, making it the fastest growing medical scheme in South Africa and the country's second largest.

The introduction of GEMS allows us to estimate how the extension of health care coverage to uninsured public sector employees has impacted health care utilisation and provider choice. The South African health system is an interesting case study to consider because of the polarized provider landscape. Following the extreme contours of one of the world's most unequal societies, there are dramatic disparities between the health services available to the poor and the affluent. Private providers tend to charge prices that are prohibitively expensive for most South Africans and consequently only a small subsection of affluent individuals (often with comprehensive coverage) has reliable and frequent access to these providers. While the private sector represents 52% of South Africa's health expenditure, only 17% of South Africans are medical scheme members. By contrast, public sector providers are visited almost exclusively by the uninsured and less affluent segment of South Africa's population.

2. Previous studies on insurance

Insurance experiments in developing countries have often been disappointing from a public health perspective, showing lower out-of-pocket expenditure and significant shifts from uncovered to covered healthcare providers, but flat utilisation rates. A randomised control trial on the impact of Seguro Popular in Mexico on utilisation and expenditure finds no impact on utilisation or on diagnoses over a 10 month period but does find a reduction in out-of-pocket expenditure on health (King et al., 2010). The absence of an impact on utilisation is ascribed to the short study period. Similarly, health insurance cover for the poor in Vietnam successfully reduced out-of-pocket expenditure on health services but did not impact total utilisation (Wagstaff, 2010). In rural Cambodia randomised allocation of insurance cover did lead to changes in provider choice (Levine et al, 2014). The authors identified an increase in the use of covered public health facilities for emergency health conditions, while the insured decreased their use of non-covered private health care and pharmacies (Levine et al, 2014). An insurance lottery in Nicaragua showed no impact on overall utilisation but shifts towards the private and public facilities covered by the insurance and a decrease in out-of-pocket expenditure (Thornton et al., 2010).

Due to the polarized provider market, the United States may be the most comparable to the South African situation. Two recent experiments show significant impact on utilisation rates. Finkelstein et al (2012) offered insurance coverage to randomly selected members of a group of low-income uninsured adults in Oregon and found a notable and significant increase in out-patient and in-patient healthcare utilization and a decrease in out-of-pocket healthcare expenditure and medical debt.

Card, Dobkin and Maestas (2008) find that eligibility for Medicare at the age of 65 leads to an increase in health care utilisation. As expected, the use of lower cost services, e.g. doctor's visits, was concentrated amongst the elderly who had the lowest rates of insurance coverage before Medicare eligibility, while higher cost and more elective type of procedures, e.g. bypass surgery and knee replacement, are mostly used by the elderly who are more likely to have held supplementary health insurance after Medicare eligibility.

Recent research in Philippines show that where health insurance is offered on a voluntary basis, there is often a lag in the utilisation of the product due to a learning process, or understanding of the working of insurance, that has to take place. Investigating household health decisions when a young child is hospitalised, Quimbo et al (2008) find significant under-utilisation of benefits by newly enrolled beneficiaries of PhilHealth. This effect was more pronounced at lower levels of maternal education. The authors argue that the underutilisation could be due to a lack of awareness of the benefits provided by PhilHealth.

Although there is no representative evidence on the gap in clinical quality between the private and public health sector, there are glaring differences in access to nurses, doctors and specialist (McIntyre et al, 2007). GPs are usually the entry point into the private provider system, while nurses represent the entry point for the public system. Survey analysis shows that public sector facility visitors are considerably more likely to complain about rude staff, drug stock outs and long waiting times. At private clinics waiting times ranged between 10 and 40 minutes versus 50 minutes to 3 hours at public sector clinics (Palmer, 2002). Consequently, it is not surprising that recent studies report substantial differences in the perceived quality of care offered by private and public providers. A discrete choice study on the nature of preferences for public healthcare in the Western Cape and Eastern Cape provinces of South Africa found "a general preference not to utilize public health facilities" and concluded that if government wanted to boost clinic visits, they would have to improve public health facilities (Honda et al, 2014: 9). Similarly, McIntyre et al (2009:725) indicate that individuals were only willing to contribute to public health services if they were assured of the quality of such health services.

Physical access and affordability are not significant constraints to demand. Public primary care services are free and the enforcement of hospital fee payments is weak and variable. The General Household Survey shows that in 2008 only 3.8% of the bottom quintile said they did not consult a health worker because the facility was too far (Burger et al, 2012). Consequently, in the South Africa system the value of health insurance lies primarily in it providing a gateway to better quality health care services. This was also the conclusion of McLaren, Ardington and Leibbrandt (2013: 11) who argued that health insurance play an "important mediating role…in accessing higher quality private care". In this way the polarization of the health provider landscape and specifically the large quality differentials between public and private providers represent a rare opportunity to investigate the effect of an exogenous expansion of insurance eligibility and subsidies in a health care system with highly variable service quality.

The research question is also of pragmatic value to local planning needs. In a South African context this question is relevant and pertinent because it provides some indication of the magnitude of the demand shifts that can be anticipated if the government contracts in the services of general practitioners or doctors from the private sector under the proposed National Health Insurance (NHI) plan.[1] Interpreting public sector employment as an exogenously administered expansion of eligibility for insurance, our estimates provide a reliable benchmark of how such a move may affect choices amongst providers, especially at the high volume entry level access points, i.e. public sector clinics and private sector GPs. Providing an estimate of the size of such shifts can help to guide the government's workforce planning and inform their rationing and gatekeeping strategies.

3. South African health system

Although there is no representative evidence on the gap in clinical quality between the private and public health sector, there are glaring differences in access to nurses, doctors and specialist (McIntyre et al, 2007). GPs are usually the entry point into the private provider system, while nurses represent the entry point for the public system. Survey analysis shows that public sector facility visitors are considerably more likely to complain about rude staff, drug stock outs and long waiting times. At private clinics waiting times ranged between 10 and 40 minutes versus 50 minutes to 3 hours at public sector clinics (Palmer, 2002). Consequently, it is not surprising that recent studies report substantial differences in the perceived quality of care offered by private and public providers. A discrete choice study on the nature of preferences for public healthcare in the Western Cape and Eastern Cape provinces of South Africa found "a general preference not to utilize public health facilities" and concluded that if government wanted to boost clinic visits, they would have to improve public health facilities (Honda et al, 2014: 9). Similarly, McIntyre et al (2009:725) indicate that individuals were only willing to contribute to public health services if they were assured of the quality of such health services.

Physical access and affordability are not significant constraints to demand. Public primary care services are free and the enforcement of hospital fee payments is weak and variable. The General Household Survey shows that in 2008 only 3.8% of the bottom quintile said they did not consult a health worker because the facility was too far (Burger et al, 2012). Consequently, in the South Africa system the value of health insurance lies primarily in it providing a gateway to better quality health care services. This was also the conclusion of McLaren, Ardington and Leibbrandt (2013: 11) who argued that health insurance play an "important mediating role…in accessing higher quality private care". In this way the polarization of the health provider landscape and specifically the large quality differentials between public and private providers represent a rare opportunity to investigate the effect of an exogenous expansion of insurance eligibility and subsidies in a health care system with highly variable service quality.

[1] The South African government released a policy proposal on the implementation of a NHI scheme in 2011 (Government of South Africa, 2011).

The research question is also of pragmatic value to local planning needs. In a South African context this question is relevant and pertinent because it provides some indication of the magnitude of the demand shifts that can be anticipated under the proposed National Health Insurance (NHI) plan. The South African government released a policy proposal on the implementation of a NHI scheme in 2011 (Government of South Africa, 2011).[2] In the NHI policy proposal, it is suggested that the NHI will contract in the services of general practitioners or doctors from the private sector (Government of South Africa, 2011). Interpreting public sector employment as an exogenously administered expansion of eligibility for insurance, our estimates provide a reliable benchmark of how the proposed NHI may affect choices amongst providers, especially at the high volume entry level access points, i.e. public sector clinics and private sector GPs. Providing an estimate of the size of such shifts can help to guide the government's workforce planning and inform their rationing and gatekeeping strategies.

4. Government Employee Medical Scheme

The 1999 Remuneration Policy Review by the South African government identified a number of concerns with health insurance provision, including "inequality in access to medical scheme cover, affordability concerns, lack of value for money, spending inefficiencies and little integration with public sector health care" (McLeod and Ramjee, 2007: 55). At the time less than half of employees had health insurance cover (McLeod & Ramjee, 2007). [3]

In response to these problems, a legal framework was drafted for the establishment of a government employees medical scheme in 2002 (McLeod & Ramjee, 2007), following which GEMS was legally registered in January 2005 and started to actively recruit members in January 2006 (GEMS, 2012). To encourage government employees enrolled in open schemes to join the newly established health insurance scheme in the absence of a track record on good administration and coverage, benefit options were competitively priced[4] and the government offered increased subsidies of 75% (cf. 66% for open schemes) to workers enrolled in GEMS.[5] The government provided a 100% subsidy of the lowest tier benefit option to employees on salary level 1-5 (equivalent to a threshold of R9000 per month in 2013). [6]

[2] If fully implemented, the NHI will provide universal health coverage to all South Africans. Although no further details on the financing component of the proposal have since been publically released, government officials have continued to refer to NHI as the government's position on the future financing of healthcare in South Africa (Gordan, 2014; Motsoaledi, 2014).

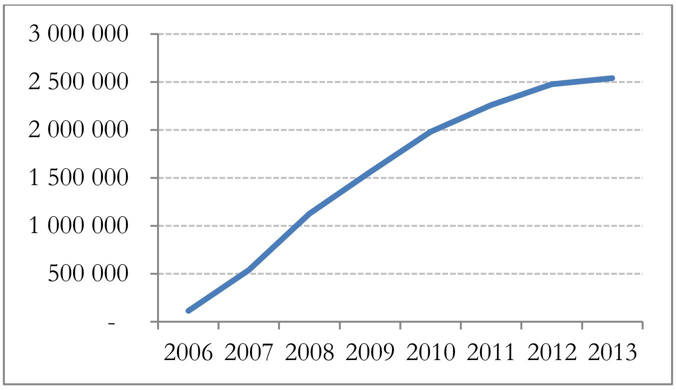
[3] It is, however, important to note that despite the criticism against the previous public sector approach to health insurance, at the time GEMS was launched government employees were already receiving quite generous health insurance subsidies. Government provided a subsidy of up to two-thirds of the cost of any health insurance scheme selected by staff members (McLeod & Ramjee, 2007).

[4] According to GEMS own comparisons outlined in their most recent online fact sheet (Available as part of their media package from the following link: https://goo.gl/gF36N3) benefits were provided at a discount of 10% to 25% cf. other health insurance.

[5] To limit the government's exposure to increasing insurance premiums, a cap was proposed for the subsidy (McLeod & Ramjee, 2007). Initially the cap was R2020 but by 2013 the subsidy limit for Sapphire had increased to R2760 for these workers (GEMS, 2013).

[6] GEMS offer five benefit options: Sapphire, Beryl, Ruby, Emerald and Onyx. The Sapphire and Beryl schemes are targeted at lower income employees. Sapphire beneficiaries were restricted to public hospitals and Beryl beneficiaries were offered access to a network of private hospitals. Ruby, Emerald and Onyx were more generous and covered hospitalisation in any private hospital. The Ruby option includes a personal medical savings account which can be used for day-to-day medical expenses, while the Emerald and Onyx options offer comprehensive health coverage for both private out-patient and in-patient providers and are mainly targeted at higher income earners. Emerald is the largest scheme and has been designed to be comparable to health insurance packages offered by open schemes.

Figure 1: Total lives (members and beneficiaries) covered by GEMS, 2006 - 2013



Source: Total members and beneficiaries on 31 December of each year from GEMS annual reports 2006-2013.

Figure 1 shows steady growth in enrolment in GEMS between 2006 and 2013. By 31 December 2013, GEMS covered more than 2.5 million lives via 684,281 government employees. These 2.5 million lives represented 5% of the South African population and more than a quarter of the South African health insurance market. More than half of these members were previously uninsured (Moroabi, 2013). By the end of 2013, 60.3% of public sector employees were members of GEMS (GEMS, 2013). [7]

5. Methodology

5.1 Identification strategy

Our econometric model considers health seeking behaviour, y, of individual i at period t. This is expressed as a function of whether or not the individual has health insurance (in which case $h = 1$, otherwise $h = 0$), as well as other observable ($\boldsymbol{x}$) and unobservable (u) determinants

$$y_{it} = \alpha h_{it} + \boldsymbol{x}_{it}\boldsymbol{\beta} + u_{it} \qquad [1]$$

The interest of our econometric analysis is primarily in identifying and estimating the causal effect of insurance on health seeking behaviour, represented by the treatment effect parameter α.

Medical scheme members effectively face an altered set of health service prices due to medical scheme reimbursements but also subsidies and tax benefits associated with insurance. However, identifying this effect using observational data is complicated by the fact that health insurance is not randomly administered to individuals, but is rather the outcome of an expected cost-benefit analysis based on a variety of considerations many of which are unobserved by the econometrician. We would expect individuals with more resources, who place greater importance on good health or with a history of health problems to be more inclined to seek expensive health care regardless of whether insured or not. However, these are the same individuals who would have more to gain from medical insurance coverage and hence more likely to be insured. A positive correlation between insurance and medical visits is therefore not necessarily indicative of the causal effect of insurance on health care choice. Controlling for measures of perceived health, household income and other covariates should partly address this problem, but some upward bias in the estimate of α is likely to remain due to the presence of unobservable

[7] Health insurance coverage in the public sector is, however, likely to be higher than the 60.3% as employees who were previously members of other schemes were allowed to retain their membership, albeit foregoing the more generous subsidies for GEMS.

resource constraints, preference factors and health history. On the other hand, if health insurance is measured with error then its effect on behaviour may suffer from attenuation bias.

An instrumental variable strategy can identify the causal impact of insurance, but convincing instrumental variables are notoriously rare. An instrumental variable is required to affect the inclination to have medical insurance without directly affecting health seeking behaviour. We believe the introduction of GEMS provides us with such an instrument. In 2006 all government employees became eligible for insurance subsidies, the subsidy was increased and there were significant efforts to provide attractive health insurance coverage options for employees with low salaries. Crucially, GEMS only affected the behaviour of public sector employees and their households, which suggest that other workers may – after making the necessary adjustment for differences in composition – provide a useful counterfactual for behaviour and choices in the absence of this scheme.

Our proposed identification strategy combines elements of the difference-in-difference and instrumental variable estimators. In the first-stage regression we use a difference-in-difference approach that allows us to extract the exogenous variation in health insurance due to the implementation of GEMS. Suppose health insurance is determined according to the following process

$$h_{it} = \boldsymbol{x}_{it}\boldsymbol{\theta} + \pi p_{it} + \delta_t + \gamma_t p_{it} + e_{it} \qquad [2]$$

where p_{it} is a public sector dummy variable ($p_{it} = 1$ if the individual is a government employee, and 0 otherwise), δ_t represents the period t health insurance effect and γ_t is the additional post-GEMS government employee effect. The public sector effect, π, controls for the fact that public sector employees may be different from other workers in unobservable ways that affect the health insurance choice, while the time effects δ_t represent economy-wide changes in medical costs, health awareness and other unobservable determinants of the decision to obtain health insurance. Finally, the γ_t coefficients are defined so that $\gamma_t = 0$ if $t < \tau$, where τ is the date of the implementation of GEMS. Equation [2] is estimated by regressing medical scheme coverage on the observable covariates, a set of time dummies, a public sector dummy and a public sector dummy interacted with time dummies for all periods since the implementation of GEMS in 2006. It is worth noting that our use of time dummies allows for a more general time trend than a specification that includes a linear time trend only. Furthermore, one can explicitly test the validity of the assumption that public sector workers have a similar pre-GEMS time trend as the rest of the economy by interacting the public sector dummy with all the period dummies and formally testing whether $\gamma_1 = \cdots = \gamma_{\tau-1} = 0$.

The coefficients on the time-government sector interactions represent the difference-in-difference first-stage effects of interest. Under the assumption that $E(e_{it}|p_{it}, t, \boldsymbol{x}_{it}) = 0$ it follows that

$$\begin{aligned}\gamma_\tau = \{&E(h_{it}|p_{it} = 1, t = \tau, \boldsymbol{x}_{it}) - E(h_{it}|p_{it} = 0, t = \tau, \boldsymbol{x}_{it})\} \\ &- \{E(h_{it}|p_{it} = 1, t = \tau - 1, \boldsymbol{x}_{it}) - E(h_{it}|p_{it} = 0, t = \tau - 1, \boldsymbol{x}_{it})\}\end{aligned}$$

and similarly for γ_t where $t > \tau$. Intuitively, if the time trend in health insurance coverage for those not in the public sector provides an accurate counterfactual for how government employees would have behaved in the absence of GEMS, then γ_t represents the causal effect of GEMS on the probability of having health insurance in period $t > \tau$ for government employees. These effects can be consistently estimated using OLS, and the estimates used to construct a predicted health insurance membership variable $\hat{h}_{it}$ which is purged of all endogenous variation. Rewriting equation [1] to include public sector and time dummies, and replacing health insurance with its predicted value, $\hat{h}_{it} = \boldsymbol{x}_{it}\widehat{\boldsymbol{\theta}} + \hat{\pi}p_{it} + \hat{\delta}_t + \hat{\gamma}_t p_{it}$, produces the following reduced form equation:

$$y_{it} = \alpha\hat{h}_{it} + \boldsymbol{x}_{it}\boldsymbol{\beta} + \mu p_{it} + \omega_t + v_{it} \qquad [3]$$

A least squares regression of equation [3] will now provide a consistent estimate of the treatment effect, as long as $E(u_{it}|p_{it}, t, \boldsymbol{x}_{it}, \hat{h}_{it}) = 0$. The crucial restriction is that variation obtained from our instrumental variables must be uncorrelated to the unobservable determinants of health care choice after conditioning on period, public sector employment and the vector of observable covariates $\boldsymbol{x}_{it}$. This is analogous to our assumption of a common trend in health insurance coverage for public sector employees and the rest of the population in the absence of GEMS.

5.2 Two-sample two-stage least squared estimation

The first and second stage equations of a two-stage least squares (2SLS) estimator are typically estimated using the same sample. However, following the seminal article by Angrist and Krueger (1992), many empirical studies have used a two-sample two-stage least squares (TS2SLS) approach to estimate the treatment effect of interest. This approach is necessary when no single dataset contains the instrumental variable, the treatment variable and the outcome variable, but there are two different datasets that contain the instrumental and the treatment variables, and the instrumental and outcome variables, respectively. As discussed in section 5, this is the case for the first part of our empirical analysis: one dataset contains data on industry of employment and health insurance status, while another contains data on industry of employment and health seeking behaviour.

The fact that the first and second stages are estimated with different samples does not pose a problem for the identification of the parameter of interest. The assumption that $E(e_{it}|p_{it}, t, \boldsymbol{x}_{it}) = 0$ is sufficient to ensure consistent estimates of the coefficients in equation [2], and these estimates can then be used to calculate predicted values of $\hat{h}_{it}$ in the second sample. If the assumption that $E(u_{it}|p_{it}, t, \boldsymbol{x}_{it}, \hat{h}_{it}) = 0$ is also valid, then the second stage coefficients will also be consistently estimated. Providing that both datasets are sufficiently large to invoke the law of large numbers, the fact that we are using two samples rather than one does not affect the consistency of our estimate of α.

However, the fact that the second stage regression uses a generated regressor means that the standard errors of the estimates should be adjusted to reflect the sampling variability in this variable. This is relatively simple to do for a single sample 2SLS estimator, but somewhat more complicated in the two sample case. Inoue and Solon (2010) review the different approaches to estimating the covariance matrix, and propose using the method first derived by Murphy and Topel (1985). A similarly appropriate, albeit more computationally intensive technique is to use the bootstrap approach.

5.3 Heterogeneous and marginal treatment effects

The simplest interpretation of the 2SLS estimates supposes that the effect of health insurance on medical visits is the same for everyone in the population, in which case the treatment effect is a constant parameter. In reality this effect may actually vary across individuals based on factors such as the individual's state of health, the distance to medical facilities and household income. From a choice theoretic perspective, one would expect higher rates of health insurance coverage amongst those groups that have more to gain from being insured, and this kind of sorting on gains complicates estimation of the treatment effect.

In order to investigate this issue formally, let y_1 and y_0 be the binary variables representing whether or not the individual will visit a medical facility if insured and uninsured, respectively, and let c be the cost of insurance. We denote the binary insurance variable as d, where $d = 1$ if insured and $d = 0$ if not. The gross benefit of health insurance (expressed in terms of medical visits) is $y_1 - y_0$, while the net benefit is $y_1 - y_0 - c$. The fact that we cannot ever observe both potential health visit outcomes is the fundamental problem of causal inference.

Suppose the econometrician observes a set of covariates $\boldsymbol{x}$ that potentially affect health visits and the cost of insurance, as well as instrumental variables $\boldsymbol{z}$ that affect only insurance costs. It is sometimes convenient to express the potential treatment status for specific values of the instruments as $d(\boldsymbol{z})$. Imbens and Angrist (1994) introduce the following three assumptions: 1) the potential health visits (y_0, y_1) and $d(z)$ are independent of $\boldsymbol{z}|\boldsymbol{x}$, 2) $P(d = 1|\boldsymbol{x}, \boldsymbol{z})$ is a non-trivial function of $\boldsymbol{z}|\boldsymbol{x}$, and 3) for any two values of z, either $d(z^1) \geq d(z^2)$ or $d(z^1) \leq d(z^2)$ for all individuals. Under these assumptions the 2SLS estimator can be shown to estimate the local average treatment effect (LATE). In the case of our model and choice of instrument, this represents the average effect of health insurance on the probability of seeking medical treatment for individuals whose insurance choice was affected by the implementation of GEMS.

One drawback of the LATE interpretation is that the group of compliers need not be a representative subsample of the population, in which case the LATE may provide misleading predictions of the effects of other, as yet unobserved policies. This concern certainly applies to our use of GEMS, which only affected public sector workers who did not already have insurance before the implementation of GEMS. However, Heckman and Vytlacil (1999, 2005) demonstrate that the LATE assumptions can also be used to estimate the marginal treatment effect (MTE) for various other groups that were unaffected by the variation in the instrumental variables. This approach could potentially allow us to extrapolate our estimated effects to parts of the population not yet covered by health insurance.

The Heckman and Vytlacil approach requires considering the behaviour of an economic agent who makes rational decisions in a context of imperfect information about the future. Suppose that an individual chooses whether or not to obtain insurance based on the net benefit they expect to derive from insurance:

$$i_d = E(y_1 - y_0 - c|\Omega)$$

where Ω is the information set at the time of making the insurance decision. Furthermore, suppose the health outcome and cost variables can be expressed as linear functions of the observable determinants of the medical visit, $\boldsymbol{x}$, and insurance costs $\boldsymbol{z}$:

$$y_0 = \boldsymbol{x}\boldsymbol{\beta}_0 + u_0$$

$$y_1 = \boldsymbol{x}\boldsymbol{\beta}_1 + u_1$$

$$c = \boldsymbol{x}\boldsymbol{\beta}_c + \boldsymbol{z}\boldsymbol{\pi} + u_c$$

where $E(y_d|\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}_d$ and $E(c|\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}\boldsymbol{\beta}_c + \boldsymbol{z}\boldsymbol{\pi}$. It then follows that

$$i_d = \boldsymbol{x}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_c) - \boldsymbol{z}\boldsymbol{\pi} + E(u_1 - u_0 - u_c|\Omega) = \mu_d(\boldsymbol{x}, \boldsymbol{z}) - v$$

where $\mu_d(\boldsymbol{z}) \equiv \boldsymbol{x}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_c) - \boldsymbol{z\pi}$ and $v \equiv -E(u_1 - u_0 - u_c|\Omega)$. The individual's health insurance choice can then be expressed as

$$d = 1(i_d > 0) = 1(\mu_d(\boldsymbol{x}, \boldsymbol{z}) > v)$$

Let be $F_v(.)$ be the strictly increasing cumulative distribution function of v. From the perspective of the econometrician, who can observe $(\boldsymbol{x}, \boldsymbol{z})$ but not (u_1, u_0, u_c) or v, the conditional probability of having insurance is

$$p(\boldsymbol{x}, \boldsymbol{z}) = P(d = 1|\boldsymbol{x}, \boldsymbol{z}) = P(\mu_d(\boldsymbol{x}, \boldsymbol{z}) > v) = F_V\big(\mu_d(\boldsymbol{x}, \boldsymbol{z})\big)$$

This allows us to define a new variable $u_d \equiv F_v(v)$ which is distributed uniformly over the unit interval. The values of u_d represents the quantiles of the unobservable net benefit of health insurance, v. High values of u_d correspond to low values of $E(u_1 - u_0 - u_c|\Omega)$, e.g. a small unobservable effect of insurance of medical visits or a high unobservable insurance cost. Such individuals will only purchase insurance if their observable covariates are such that $p(\boldsymbol{x}, \boldsymbol{z})$, the conditional probability of being insured is high enough to exceed u_d. The LATE assumptions ensure that u_d is independently distributed of $\boldsymbol{z}|\boldsymbol{x}$. Heckman (2010) shows that the LATE estimator obtained with a binary instrumental variable $z \in \{z^1, z^2\}$ can be expressed in terms of the values of this unobserved variable

$$LATE(z^1, z^2) = E\big(y_1 - y_0|p(z^1) \leq u_d \leq p(z^2)\big)$$

In other words, it is the average gross benefit for individuals with values of u_d between $p(z^1)$ and $p(z^2)$.

Vytlacil (2002) demonstrates that the LATE assumptions also imply that the instrumental variables only affect the conditional expectation of the observable outcome $y = y_0 + (y_1 - y_0)d$ through its effect on the propensity score $p(\boldsymbol{x}, \boldsymbol{z})$: $E(y|\boldsymbol{x}, \boldsymbol{z}) = E\big(y|\boldsymbol{x}, p(\boldsymbol{x}, \boldsymbol{z})\big)$. This is a particularly useful property when working with a multivalued instrument (or instrument vector). It follows that

$$\begin{aligned} E(y|\boldsymbol{x}, \boldsymbol{z}) = E\big(y|\boldsymbol{x}, p(\boldsymbol{x}, \boldsymbol{z})\big) &= E\big(y_0|\boldsymbol{x}, p(\boldsymbol{x}, \boldsymbol{z})\big) + pE(y_1 - y_0|\boldsymbol{x}, p(\boldsymbol{x}, \boldsymbol{z}), d = 1) \\ &= \boldsymbol{x\beta}_0 + pE(y_1 - y_0|\boldsymbol{x}, p(\boldsymbol{x}, \boldsymbol{z}), d = 1) \end{aligned} \quad [4]$$

The MTE can now be defined as

$$MTE(\boldsymbol{x}, p) = \frac{\partial E(y|\boldsymbol{x}, \boldsymbol{z})}{\partial p} = E\big(y_1 - y_0|\boldsymbol{x}, p(\boldsymbol{x}, \boldsymbol{z})\big)$$

Intuitively, this is the average effect of a marginal change in the probability of having health insurance on the probability of visiting a clinic for individuals with $(\boldsymbol{x}, p)$. By decreasing the cost of insurance and thereby increasing the conditional probability of having insurance $p(\boldsymbol{x}, \boldsymbol{z})$ of certain workers, the implementation of GEMS allows us to identify the effect of insurance on the health seeking behaviour of workers with values of u_d that were close to the affected values of $p(\boldsymbol{x}, \boldsymbol{z})$.

Since the left-hand side of equation [4] can be consistently estimated with sample data, it is possible to identify the MTE by observing how health seeking behaviour changes when some individuals are induced into health insurance due to exogenous variation in $p(\boldsymbol{x}, \boldsymbol{z})$. It is worth pointing out that, somewhat counter-intuitively, individuals with high levels of $p(\boldsymbol{x}, \boldsymbol{z})$ are used to identify the marginal treatment effect for individuals with high values of u_d, i.e. those who have a low unobservable net insurance benefit.

Equation [4] can be estimated using semi-parametric techniques but, as Carneiro et al (2011: 2761) point out, conditioning on $\boldsymbol{x}$ nonparametrically can be very difficult when there are many covariates. An alternative approach is to replace the LATE assumption that the potential outcomes are independent of $\boldsymbol{z}|\boldsymbol{x}$ with the stronger assumption that these outcomes are independent of $(\boldsymbol{x}, \boldsymbol{z})$. In this case we can estimate equation [4] by regressing observed medical visits on the list of control variables and a flexible function of the estimated value of $p(\boldsymbol{x}, \boldsymbol{z})$, possibly a low order polynomial.

6. **Data**

The estimation strategy outlined above requires information on health seeking behaviour, health insurance and the industry of employment from 2002 to 2012. Unfortunately, no single data set meets all these criteria, so we pursue two alternative estimation strategies. The first utilises the reliable health insurance and industry information in the labour force surveys (the biannual Labour Force Survey, or LFS, from 2000 to 2007 and the Quarterly Labour Force Surveys, or QLFS, since 2008)[8] and the useful and detailed information on health services in the general household surveys (GHS)[9]. This approach is required because the LFSs do not have any information on health seeking behaviour and the GHSs undercaptures medical scheme membership during the crucial first years of GEMS (Figure 2). These two datasets are then combined to produce a two-sample 2SLS estimate of the causal effect of interest. The first-stage estimates equation [2] using the LFS/QLFS data, after which these estimates are applied to the GHS data in order to estimate equation [3] with the instrumented insurance variable.

The second strategy uses the three waves (2008, 2010, 2012) of the National Income Dynamics Study to estimate the first and the second stage of the instrumented variable regression. The National Income Dynamics Survey is a nationally representative panel survey covering about 7000 households. Because the first wave occurs in 2008, the data does not allow comparison with the pre-GEMS period and we therefore use the gradual GEMS roll out (see Figure 1) and consequent increase in the likelihood of being a member of a medical scheme amongst public sector workers over the four-year window to capture the exogenous impact of GEMS.

Replicating estimates from one data source on another is always a useful way to gauge the reliability of the estimates, but there are a number of differences between the GHS/LFS/QLFS data and the NIDS data that makes a comparison of these estimates particularly useful. First, the GHS's reporting of health visits is only asked to the subsample that has recently been ill or injured. In contrast, NIDS does not ask only about recent visits, but rather enquires about when your last visit had occurred. The filtering via a question on acute illness and injury also means that the GHS data is likely to reflect mainly visits for acute illness and serious cases of chronic illness, whereas NIDS data may more accurately reflect other dimensions of health seeking behaviour like antenatal and preventative care as well as less severe cases of chronic illness. Furthermore, the NIDS data period is from 2008 until 2012 and thus starts two years after the launch of GEMS, which may be advantageous given the slow initial enrolment rate shown in Figure 1. Furthermore, the NIDS data also contains a richer set of covariates, including whether the individual experienced a number of health symptoms and these variables can be used as control variables to assess the robustness of our estimates.

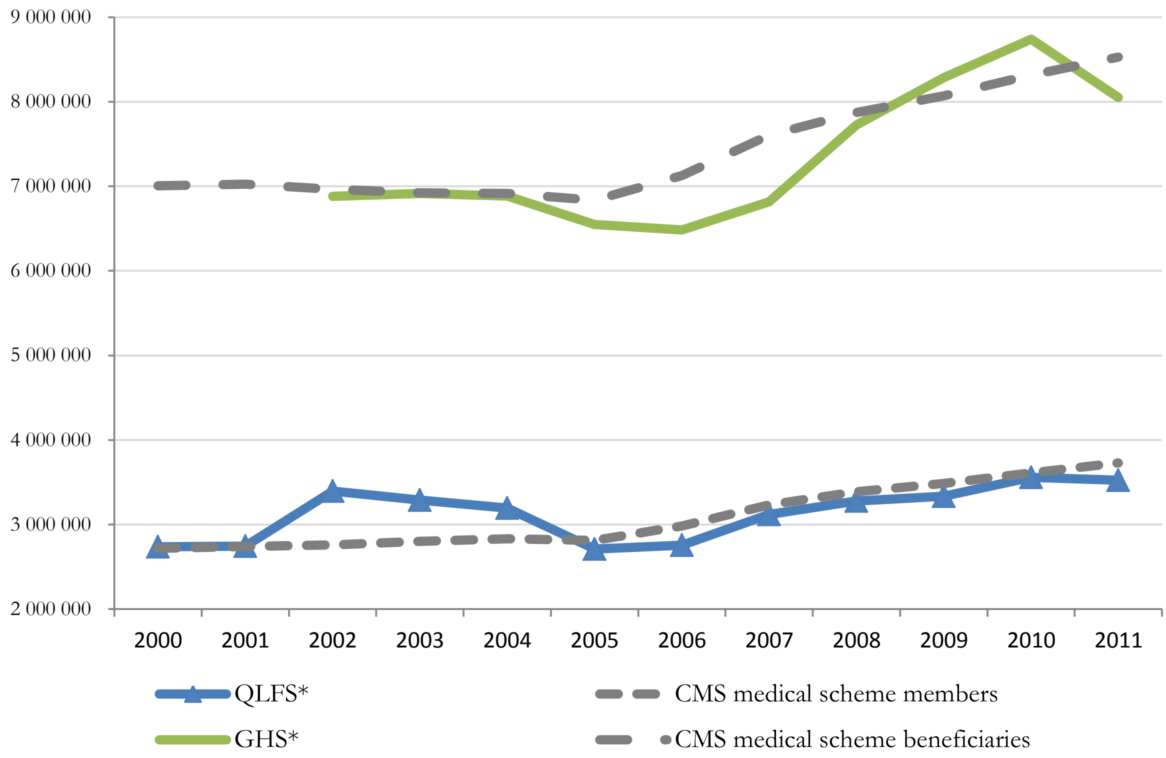
---

[8] The LFS/QLFS collects information on the labour market status and activities of a sample of individuals living in South Africa who are older than 14. It covers about 30 000 dwellings.

[9] The GHS is an annual survey and aims to gather information on the circumstances and quality of life of households. It includes approximately 25 000 households.

Figure 2 below considers levels and time trends for beneficiaries and members, comparing the administrative data from the Council for Health insurance with estimates from the General Household Survey and the Labour Force Surveys. It is encouraging to see that the survey data tracks the administrative data remarkably well – the only exceptions being a slight LFS overcount between 2002 and 2004 and a GHS undercount between 2006 and 2008.

Figure 2: Growth in medical scheme members and beneficiaries, 2002-2011



Source: GHS 2002-2011 & LFS 2002-2007, QLFS 2008-2011, CMS 2002-2011

Although our empirical strategy outlined in section 4 requires identifying public sector workers from the survey data, both the GHS and the NIDS data only provide industry of employment information at the one-digit ISO level. The labour force surveys (which provide more detailed industry data, as well as sector of employment) reveal that there is substantial overlap between public sector employment and working in the community and social services (CSS) industry. This suggests using being employed in this industry as a proxy for government employment. Of course, this adds measurement error to our instrumental variables, which may affect our estimates.

Although there are some (11%) public sector workers who work outside of this industry, the more serious issue is that only 63% of workers in this industry are public sector workers. Our identification strategy assumes that there is a common time trend in health insurance and health seeking behaviour that affects all non-government workers, including the non-government workers in the CSS industry. This implies that the OLS estimates of coefficient γ_2 in equation [2] will be attenuated by 0.63, since this coefficient is the weighted average of the true GEMS effect on government employees and zero on the remainder of the industry. Similarly, regressing health seeking behaviour on the interaction between the CSS industry and the post-GEMS periods will suffer from the same attenuation rate. Since the 2SLS estimator can be calculated as the ratio of these two effects, the attenuation effects will cancel out to produce a consistent estimate of the causal effect of interest, even where some non-government workers are included in our CSS industry variable.

The panel structure of the NIDS data was exploited to obtain a cleaner version of the public sector variable. Particularly, individuals who claimed to have transitioned between sectors with no corresponding variation in their job tenure variable are reclassified in order to produce internally consistent responses to the industry of employment and job tenure questions.

7. Empirical analysis

7.1 The effect of insurance on utilisation

We begin our analysis by investigating the trends in medical scheme coverage for government employees and non-government employees between 2003 and 2008 using the LFS-QLFS data. The GHS ceased to gather industry data for a number of years after 2008 and thus our analysis is limited to a three-year window shortly after the introduction of GEMS in 2006. Figure 3 reveals that public sector employees are substantially more likely than other working aged South Africans to have health insurance in all periods. Furthermore, we can observe that trends for the two groups are very similar before 2006 but appear to diverge thereafter, with government employees experiencing a more rapid increase in their likelihood of having medical insurance. In fact, a formal test of the hypothesis of a common trends is not rejected for the 2003-2005 years (with a p-value of 0.3464), but is rejected for the GEMS years of 2006-2008 (with a p-value of less than 0.0001 and an F-test statistic of 22). The very similar pre-implementation trends experienced by government employees and other working aged South Africans provides support for our identification strategy of assuming that the time trend for non-government employees offers a counterfactual for what would have happened to government workers in the absence of GEMS.

The results from the two-sample 2SLS analysis using the LFS/QLFS and the GHS are shown in Table 1 below. The first-stage regression results obtained without any controls (apart from year dummies and a public sector dummy) in column 1 show that there is a positive and significant coefficient on the 2008 year interaction with government sector, validating the IV approach. In fact, the first-stage coefficients indicate that by 2008 GEMS had increased the proportion of public sector employees with health insurance from 57% to 60%. The associated F-test statistic is large, dispelling any concerns regarding weak instruments. According to the second stage estimates in column 2 membership of health insurance increases the likelihood that ill individuals will consult a health worker by 77%. However, when using the bootstrapped standard errors[10] this effect is not statistically significant.

[10] All bootstrapped standard errors are calculated with 50 repetitions.

Figure 3: Share of working age individuals with health insurance, 2003-2008

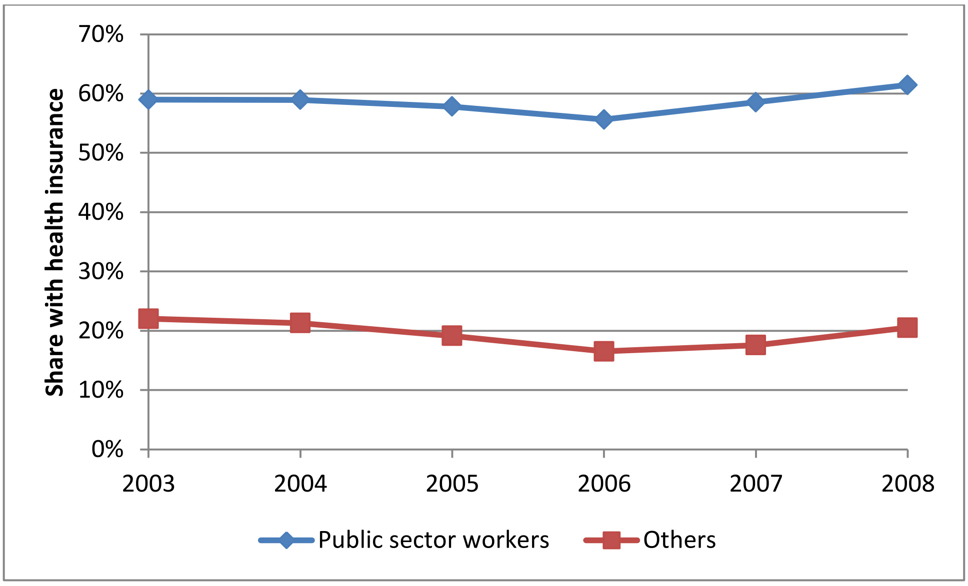


Table 1: Impact of insurance on utilisation, 2003 – 2008 (LFS/QLFS, GHS)

VARIABLES	(1) Health insurance	(2) Health visit if ill	(3) Health insurance	(4) Health visit if ill
Public sector	0.365***	-0.206	0.227***	-0.124
	(0.00287)	(0.204)	(0.00271)	(0.0971)
Health insurance (predicted)		0.766		0.757*
		(0.558)		(0.413)
Public sector*2006	-0.0294***		-0.0249***	
	(0.00576)		(0.00521)	
Public sector*2007	-0.000259		0.00337	
	(0.00565)		(0.00511)	
Public sector*2008	0.0287***		0.0440***	
	(0.00442)		(0.00401)	
Observations	332,532	18,937	329,692	18,774
R-squared	0.110	0.011	0.276	0.026
F-test	33.69		66.86	
Year effects	Y	Y	Y	Y
Control variables	N	N	Y	Y
Bootstrapped s.e.	N	Y	N	Y

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Impact of covariates on insurance and utilisation

VARIABLES	GHS/LFS/QLFS (1) Health insurance	(2) Health visit if ill	(3) Health visit if ill	NIDS (4) Health insurance	(5) Health visit	(6) Health visit
Health insurance (predicted)			0.757**			0.791***
			(0.320)			(0.302)
Public sector	0.227***	0.0524***	-0.124*	0.248***	0.0693***	-0.162*
	(0.00271)	(0.00777)	(0.0748)	(0.0146)	(0.0126)	(0.0895)
Public sector*2006	-0.0249***					
	(0.00521)					
Public sector*2007	0.00337					
	(0.00511)					
Public sector*2008	0.0440***					
	(0.00401)					
Public sector*2010				0.0475**		
				(0.0206)		
Public sector*2012				0.101***		
				(0.0216)		
Coloured	0.132***	0.0358***	-0.0642	0.0873***	0.0785***	0.00949
	(0.00258)	(0.0123)	(0.0440)	(0.0122)	(0.0159)	(0.0317)
Indian	0.115***	0.0137	-0.0731*	0.288***	0.166***	-0.0615
	(0.00375)	(0.0187)	(0.0412)	(0.0177)	(0.0230)	(0.0906)
White	0.188***	-0.00885	-0.151**	0.351***	0.0862***	-0.191*
	(0.00212)	(0.00936)	(0.0608)	(0.0109)	(0.0142)	(0.107)
Male	0.0671***	-0.0319***	-0.0826***	0.0435***	-0.131***	-0.166***
	(0.00134)	(0.00583)	(0.0222)	(0.00605)	(0.00787)	(0.0159)
Age	0.0207***	0.0108***	-0.00489	0.0100***	0.00507***	-0.00279
	(0.000403)	(0.00140)	(0.00676)	(0.00148)	(0.00192)	(0.00367)
Age squared	-0.000201***	-8.62e-05***	6.65e-05	-6.07e-05***	-1.77e-05	2.90e-05
	(5.05e-06)	(1.62e-05)	(6.65e-05)	(1.79e-05)	(2.33e-05)	(3.12e-05)
Years of schooling	0.0231***	0.00579***	-0.0117	0.0259***	0.00611***	-0.0144*
	(0.000223)	(0.000906)	(0.00745)	(0.000995)	(0.00129)	(0.00796)
Skilled occupation	0.0459***	-0.0143**	-0.0489***	0.0883***	0.00469	-0.0647**
	(0.00157)	(0.00691)	(0.0162)	(0.00720)	(0.00936)	(0.0284)
Highly skilled occupation	0.184***	-0.0250***	-0.164***	0.188***	-6.23e-05	-0.149**
	(0.00221)	(0.00965)	(0.0597)	(0.00951)	(0.0124)	(0.0584)
Eastern Cape	0.0656***	0.0430***	-0.00686	0.0454***	-0.0551***	-0.0915***
	(0.00288)	(0.0134)	(0.0250)	(0.0144)	(0.0187)	(0.0248)
Northern Cape	0.0379***	0.0223	-0.00746	0.0396*	0.00729	-0.0236
	(0.00477)	(0.0221)	(0.0254)	(0.0208)	(0.0270)	(0.0319)
Free state	0.0696***	0.0523***	-0.000770	0.0926***	0.0136	-0.0604*
	(0.00334)	(0.0142)	(0.0266)	(0.0161)	(0.0209)	(0.0364)
KwaZulu-Natal	0.0492***	0.0167	-0.0205	0.0176	-0.0415**	-0.0557***
	(0.00271)	(0.0133)	(0.0206)	(0.0130)	(0.0169)	(0.0194)
Northwest	0.128***	0.0261*	-0.0705	0.150***	-0.0232	-0.142***
	(0.00331)	(0.0143)	(0.0433)	(0.0156)	(0.0203)	(0.0507)
Gauteng	0.0828***	0.0312***	-0.0317	0.0733***	-0.0330**	-0.0915***
	(0.00248)	(0.0115)	(0.0290)	(0.0117)	(0.0152)	(0.0279)
Mpumalanga	0.0945***	0.0271*	-0.0446	0.0784***	-0.0527***	-0.115***
	(0.00338)	(0.0143)	(0.0335)	(0.0148)	(0.0193)	(0.0319)
Limpopo	0.0871***	0.0493***	-0.0171	0.00904	-0.108***	-0.116***
	(0.00343)	(0.0159)	(0.0322)	(0.0151)	(0.0197)	(0.0217)
Symptoms				-0.00833	0.218***	0.224***
				(0.00620)	(0.00806)	(0.00915)
Coughed blood				0.0283	0.0636*	0.0405
				(0.0255)	(0.0331)	(0.0373)

Vomiting				0.0670***	0.196***	0.144***
				(0.0245)	(0.0318)	(0.0402)
Diarrhoea				-0.0321*	0.0708***	0.0951***
				(0.0185)	(0.0240)	(0.0279)
Observations	329,692	18,774	18,774	14,755	14,755	14,755
R-squared	0.276	0.025	0.026	0.358	0.118	
Year effects	Y	Y	Y	Y	Y	Y
F-test	66.86			11.11		

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Columns 3 and 4 in Table 1 also report the two-sample 2SLS estimates while controlling for possible confounding factors: demographic attributes (age, race and gender), year of schooling, province of residence and the skill level of occupation. The addition of control variables does not substantially change any of the coefficient point estimates, but it does increase the precision of the estimates. The first-stage F-statistic is larger than before and the estimated treatment effect of 76% is now statistically significant .

A closer inspection of the first-stage coefficient estimates reveals a surprising result: there occurred a temporary decrease in the share of public sector workers who reported having medical scheme coverage in 2006. The administrative GEMS data in Figure 1 demonstrated that take-up of this scheme was very low in 2006 and substantial enrolment only occurred from 2007 onwards. It is possible that during the implementation of GEMS some public sector workers perceived themselves (correctly or incorrectly) to be in a transitory phase between their old and new medical insurance schemes. We therefore perform two robustness checks in order to confirm that our results are not driven by this anomaly. First, we re-estimate the two-sample two-stage least squares model on the full sample, but now under the assumption that GEMS was only implemented in 2007. These estimates are reported in columns 1 and 2 of Table A1 in the appendix. We also re-estimate the model after omitting 2006 from the sample; the estimates are reported in columns 3 and 4 of Table A1. The results are very similar to those obtained while including 2006 as the first year of the treatment period, so we conclude that the temporary reduction in public sector health insurance in 2006 does not affect our results.

In Table A2 in the appendix we also report the results from three additional tests of the robustness of our estimates and the validity of our identifying assumptions. Columns 1 and 2 replicate the results from Table 1 estimated over a longer period that now stretches back to 2000. Although the estimated treatment effect is now somewhat smaller, the essential results are seen not to be sensitive to the choice of sample period. Perhaps the main concern with most difference-in-difference analyses is the identification of a valid control group. Individuals working in the public sector are likely to be very different from those working in, for example, the agriculture or mining industries, so we re-estimate the model with a smaller control group consisting only of employees in manufacturing, construction, wholesale and retail and financial services. These estimates (reported in columns 3 and 4) are very similar to what was obtained in our preferred specification in Table 1. Finally, columns 5 and 6 contain the results for a placebo intervention in which we estimate the effect of being in the transport, storage and communication industry during the GEMS years (relative to all other industries excepting the public sector). We observe that the first stage regression produces a very small F-test statistic, while the second stage treatment effect estimate is highly insignificant.

Our identifying assumption supposes that the only change that occurred for public sector employees with the implementation of GEMS was a change in their inclination to have health insurance. However, we may be concerned that a policy that decreased the cost of public sector health insurance also changed the selection into public sector employment. Although we cannot comprehensively test this hypothesis, we can check whether there was any change in the selection into public sector employment based on observable covariates. Table A3 in the appendix performs a 2SLS regression on the variables that are used as controls (race, age, education and occupation) in Table 1. We find no evidence that GEMS resulted in any change in selection into the public sector. Given the relatively high wages paid by the South African public sector and the high level of open unemployment, it is perhaps not surprising that the already strong selection into government employment is not substantially altered by this policy.

Table 2 reports the complete first and second stage estimates of the 2SLS regression on the LFS/QLFS/GHS data for 2003 to 2008 (in columns 1 and 3), as well as the estimates for the same regression on the NIDS data for 2008 to 2012 (in columns 4 and 6). For the purpose of comparison, we also include the estimates obtained from regressing the medical visit variable on all of the control variables only. The first-stage estimates on the NIDS data demonstrate that public sector employees continued gradually increasing their health care coverage between 2008 and 2012. The estimated effect of insurance on health care utilisation is very similar to what was obtained from the LFS-QLFS-GHS data. This suggests that this result is not sensitive to the exact wording of the question, sampling issues or whether or not we control for self-reported health measures[11].

With regard to the control variables, we observe a reassuring consistency between the estimates obtained from the LFS-QLFS-GHS data on the one hand, and the NIDS data on the other: The pattern of significance and the coefficient signs are the remarkably similar across the two sets of estimates.

7.2 The effect of insurance on provider choice

The international empirical evidence is consistent with the hypothesis health insurance only induces a large increase in health care utilisation if the health system is polarized. The 2SLS estimate in the preceding, which indicates that having health insurance causes a large increase in the inclination of South Africans to utilise health care, provides further evidence in favour of this hypothesis. Of course, if a polarized health care system is the crucial mediating factor in this relationship, then we would expect to see the increase in utilisation occur mainly due to a large increase in the utilisation of high quality health care facilities. We now investigate whether this is indeed the case.

Although both the GHS and NIDS datasets asked questions about the type of facility visited, the larger sample in the NIDS dataset is crucial in obtaining precise estimates of the effect of health insurance on facility choice[12]. These results are reported in Table 3. These estimates demonstrate that the increased use of heath care services associated with having health care is mainly due to a large increase in the likelihood of consulting a private doctor (66%) and a private hospital (12%). There is also predicted increase in public hospital use and a strong decrease in public clinic use, but neither of these effects is precisely estimated. These results provide further evidence that in countries with highly polarized health care systems, health insurance can induce a large increase in utilisation by allowing individuals to access the high cost, high quality part of the system.

[11] When not controlling for experiencing various symptoms, the 2SLS point estimate of the treatment effect is larger, but not statistically different from the estimate in column 6 of Table 2. The fact that this coefficient is larger may be indicative of larger long-term than short-term utilisation effects of insurance, and that part of the effect operates via increased awareness of poor health. However, since these differences are not statistically significant our interpretation will focus on the similarity between the estimates from the two data sources.

[12] These facility-specific effects were also estimated for the LFS-QLFS-GHS data, but apart all the estimates were statistically insignificant.

Table 3: 2SLS estimates of effect of health insurance on choice of facility (NIDS)

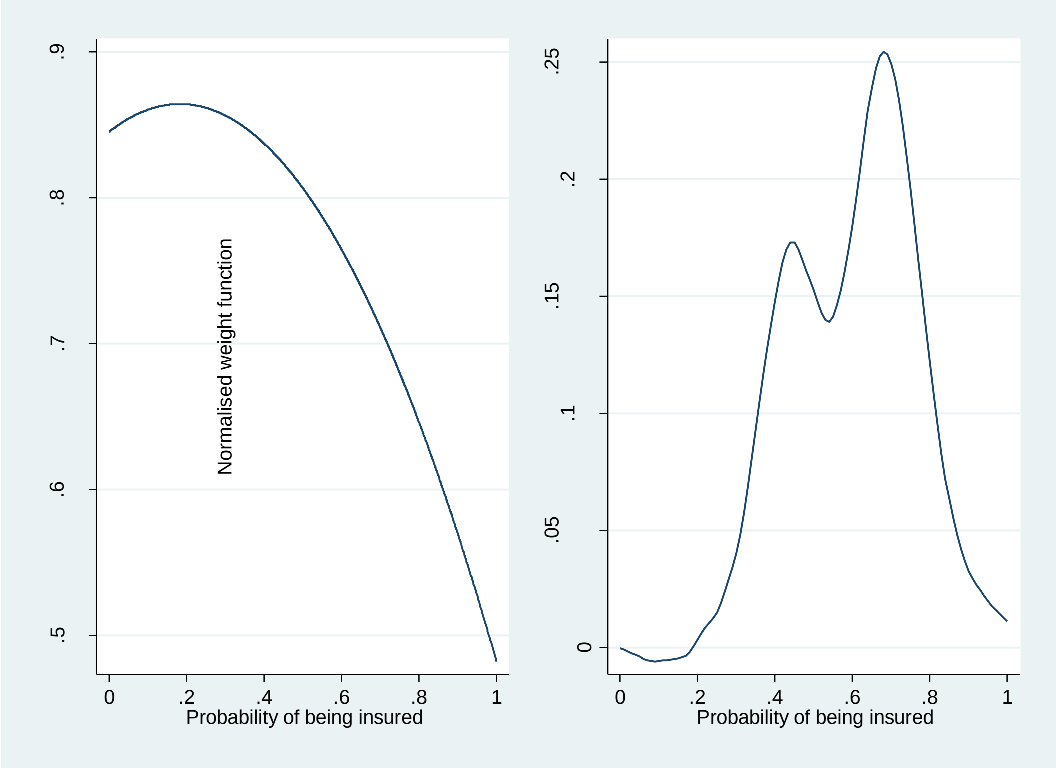
VARIABLES	(1) Any care	(2) Private doctor	(3) Private hospital	(4) Public hospital	(5) Public clinic	(6) Other
Health insurance (predicted)	0.705***	0.662***	0.121**	0.0602	-0.135	-0.00978
	(0.207)	(0.155)	(0.0550)	(0.110)	(0.151)	(0.0195)
Observations	49,990	49,980	49,980	49,980	49,980	49,980
Wave effects	Y	Y	Y	Y	Y	Y
Demographic controls	Y	Y	Y	Y	Y	Y
Education & Income	Y	Y	Y	Y	Y	Y
Symptoms	Y	Y	Y	Y	Y	Y
Geography	Y	Y	Y	Y	Y	Y

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

7.3 The marginal treatment effect of health insurance

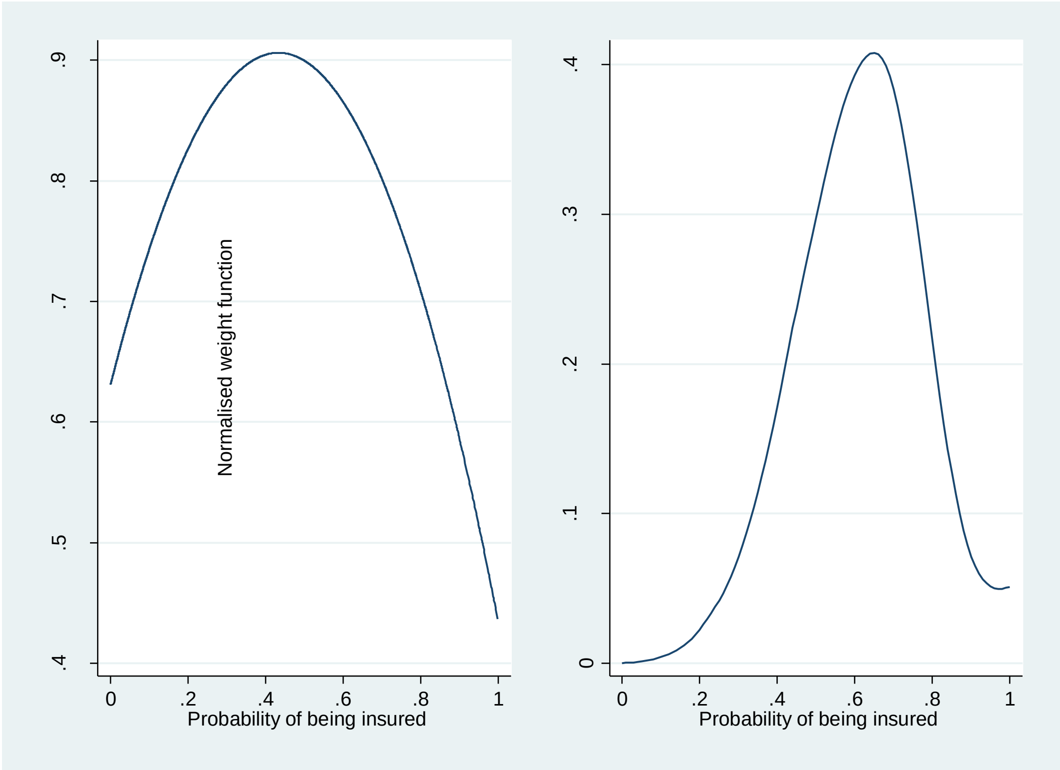
We further scrutinise the estimated effect of health insurance on utilisation by estimating the MTE for different values of the unobservable net cost of health care. In a model that allows for heterogeneous treatment effects, the 2SLS estimates reported above represent the expected treatment effect (on the probability of visiting a medical facility) for those individuals whose decision to obtain medical insurance were affected by the GEMS policy. In the model without control variables estimated in Table 1, we saw that in 2008 this group of compliers consisted of 2.87% of public sector employees who would have had a 57% probability of insurance even in the absence of GEMS. The LATE measures the weighted average of the marginal treatment effect for this small and perhaps unrepresentative group of employees.

Figure 3: Marginal treatment effect and LATE weighting function (GHS/LFS/QLFS)



Adding covariates and the stronger independence assumption referred to in section 4.3 allows the analysis to track the behaviour of individuals over a wider range of propensity scores. Our second-stage regression now includes the predicted value of the first-stage regression (the propensity score) as a third-order polynomial, which allows incremental changes in the probability of being insured to have different effects on the probability of visiting a medical facility. Plotting the first derivative of this function for different values of the propensity score produces the marginal treatment effect: the expected effect of obtaining medical insurance on the probability of visiting a medical facility for someone who induced into treatment at different propensity scores. Figures 3 and 4 show the MTE estimates for the LFS-QLFS-GHS and NIDS data respectively (in the left-hand-side panels), as well as the weighting functions that apply to the 2SLS estimates that use the implementation as GEMS as its instrumental variables. In both cases the curve is shown to be a decreasing function of the probability of having insurance in the propensity score region where most of the variation occurs. This means that the effect of insurance on medical visits is usually larger for those who are also more likely to have insurance. This type of "sorting on gains" is exactly what we would expect if individuals rationally choose whether or not to have insurance based on factors such as the availability of high quality medical facilities. It also suggests that any policy that extends the coverage of medical insurance to more households is likely to have smaller effects on medical visits. The reason why the LATEs in Tables 1 and 2 are so high is because our instruments induced exogenous variation in health care insurance amongst a group of individuals who have a higher than average marginal treatment effect.

Figure 4: Marginal treatment effect and LATE weighting function (NIDS)



8. **Conclusion**

The introduction of the GEMS created a natural experiment in the extension of health insurance eligibility. The aims of GEMS included improved equity and affordability of health insurance offered to government employees. It therefore involved an expansion of subsidy eligibility, a concerted effort to create more affordable benefit options for lower tier government workers and an increase of subsidies across the range, but also in particular for lower tier workers who qualified for a full subsidy for the most basic benefit option.

As expected in South Africa's polarized provider market, insurance causes a large response in both utilisation and provider choice. Our estimates from the two strategies show similar results and provide evidence of a very large impact of insurance cover (62 to 71%) on the utilisation of health services. Insurance increases the likelihood of using private providers and in particular private doctors (66%).

We interpret these results against the backdrop of a growing body of evidence showing strong preferences for private providers amongst South Africans. GEMS was designed partly to promote utilisation of government health facilities amongst government employees and help to generate additional revenue flows for public facilities. However, the research shows a strong shift away from government providers. Analysis of scheme data shows that even amongst the low salaried government employees, very few opted for basic benefit schemes that do not offer comprehensive private hospital cover.

While problems with the quality of public sector services are increasingly acknowledged, this analysis also clearly indicates the link between quality and quantity. Providing better quality of the health services to those without medical scheme coverage is likely to boost the per capita number of visits, which can yield significant public health benefits. For our "treated" subgroup, there appears to be few other significant constraints to demand that cannot be overcome if they can access high quality services at a low cost. It is however important to bear in mind that while this subgroup may be distinct from the traditional core of health insurance, they are employed and therefore also distinct from the lowest quintiles of South African society. Poor households are likely to face harsher trade-offs and factors such as transport costs and childcare worries may be binding constraints for this group, even when the quality of health care services has vastly improved.

The analysis also offers useful inputs for the planning process supporting the pilot and launch of National Health Insurance. The large responses in utilisation and provider choice suggest that providing access to private doctors at no cost to the user is likely to cause a large increase in total utilisation, most of which will be directed towards these private doctors. Again, the demand response of our "treated" subgroup may overestimate the response of poor South Africans that face more constraints, but in lieu of other evidence such upper bound estimates can help set parameters for scenario-based forecasts to ensure adequate workforce planning. Given the pressure on public health budgets and the lack of doctors and nurses, these estimates also highlight the need for effective and fair rationing strategies and renewed emphasis on gatekeeping to accompany other NHI health reforms.

9. References

Angrist, J. D., & Krueger, A. B., 1992. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418): 328-336.

Baicker, K., Taubman, S., Allen, H., Bernstein, M., Gruber, J.H., Newhouse, J.P., Schneider, E.C., Wright, B.J., Zaslavsky, A.M., Finkelstein, A.N. for the Oregon Health Study Group. 2013. The Oregon experiment – effects of Medicaid on clinical outcomes. *New England Journal of Medicine*. 368(18): 1713-1722.

Burger, R., Bredenkamp,C., Grobler, C. & V an der Berg, S. 2012. Have public health spending and access in South Africa become more equitable since the end of apartheid?, *Development Southern Africa,* 29:5, 681-703.

Card, D., Dobkin, C. and Maestas, N. 2008. Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare. *American Economic Review*. 98:5: 2242–2258.

Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. The American Economic Review, 101(6), 2754-2781.

Combrink, J. 2009. Is the original mandate of equitable access to affordable benefits being realised? The 2009 Annual GEMS Symposium. Adding value to the Public Service, 26 to 27 October 2009

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J.P., Allen, H., Baicker, K., and the Oregon Health Study Group. 2012. The Oregon health insurance experiment – evidence from the first year. *Quarterly Journal of Economics*. 127(3): 1057-1106.

Government Employees Medical Scheme (GEMS), 2007. Annual report. Government Employees Medical Scheme: Pretoria.

Government Employees Medical Scheme (GEMS), 2008. Annual report. Government Employees Medical Scheme: Pretoria.

Government Employees Medical Scheme (GEMS), 2009. Annual report. Government Employees Medical Scheme: Pretoria.

Government Employees Medical Scheme (GEMS), 2010. Annual report. Government Employees Medical Scheme: Pretoria.

Government Employees Medical Scheme (GEMS), 2011. Annual report. Government Employees Medical Scheme: Pretoria.

Government Employees Medical Scheme (GEMS). 2012a. GEMS and its performance. Presentation to the Portfolio Committee on Public Service and Administration. 16 May 2012.

Government Employees Medical Scheme (GEMS), 2012b. Annual report. Government Employees Medical Scheme: Pretoria.

Government Employees Medical Scheme (GEMS), 2013. Annual report. Government Employees Medical Scheme: Pretoria.

Gordan, P. 2014. Budget speech by the Minister of Finance of South Africa. 26 February. Cape Town, South Africa.

Govender, V., Chersich, M.F., Harris, Bl, Alaba O., Ataguba, J.E., Nxumalo, N. & Goudge, J. 2013. Moving towards universal coverage in South Africa? Lessons from a voluntary government insurance scheme. *Global Health Action*. 2013, 6: 19253 - http://dx.doi.org/10.3402/gha.v6i0.19253.

Government of South Africa. 2011. National health insurance in South Africa: policy paper. Pretoria, South Africa: Department of Health.

Heckman, J. J. (2010). Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy. Journal of Economic Literature, 48(2), 356-398.

Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the national Academy of Sciences, 96(8), 4730-4734.

Heckman, J. J., & Vytlacil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation1. Econometrica, 73(3), 669-738.

Honda, A., Ryan, M., Van Niekerk, R. & McIntyre, D. 2014. Improving the public health sector in South Africa: eliciting public preferences using a discrete choice experiment. *Health Policy and Planning (Advance Access)*. Doi: 10.1093/heapol/czu038.

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. Econometrica, 62(2), 467-475.

Inoue, A., & Solon, G., 2010. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3): 557-561.

King, G., Gakidou, E., Imai K. et al. 2009. Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme. *Lancet.* 373(9673): 1447–1454.

Levine, D., Polimeni, R., & Ramage, I. 2014. Insuring health or insuring wealth? An experimental evaluation of health insurance in rural Cambodia. IRLE Working Paper No. 109-14. Available at: http://irle.berkeley.edu/workingpapers/109-14.pdf (accessed on 9 September 2014).

Matsotso, M.P. & Fryatt, R. 2013. National Health Insurance: The first 18 months. In Padarath, A. & English, R. (Eds.). South African Health Review 2012/13. Durban: Health Systems Trust.

McIntyre, D., Goudge, J., Harris, B., Nxumalo, N., Nkosi, M. 2009. Prerequisites for national health insurance in South Africa: results of a national household survey. *South African Medical Journal.* 99: 725-729.

McIntyre D, Thiede M et al. 2007. A Critical Analysis Of The Current South African Health System, SHIELD Report.

McLaren, Z., Ardington, C. & Leibbrandt, M. 2013. Distance as barrier to health care access in South Africa. A Southern Africa Labour and Development Research Unit (SALDRU) Working Paper Number 97. Cape Town: SALDRU, University of Cape Town.

McLeod, H. & Ramjee, S. 2007. Health insurance. In Harrison, S. Bhana, R. & Ntuli, A. (Eds.). South African Health Review 2007. Durban: Health Systems Trust.

Moloabi, S. 2013. GEMS: Working harder to make a difference for you. 21 June 2013 www.gems.co.za (accessed on 30 October 2014)

Motsoaledi, A. 2014. Health budget vote speech by the Minister of Health. 23 July 2014. Cape Town, South Africa.

Murphy, K. M., & Topel, R. H. (1985). Estimation and Inference in Two-Step Econometric Models. Journal of Business & Economic Statistics, 3(4), 370-379. doi: 10.2307/1391724

Thornton, R.L.., Hatt, L.E., Field, E.M., Islam, M., Diaz, F.S., and González, M.A.. 2010. Social security health insurance for the informal sector in Nicaragua: a randomized evaluation. *Health Economics*. 19(S1): 181-206.

Quimbo, S., Florentino, Jhiedon, Peabody, J.W., Shimkhada, R., Panelo, C. & Solon, O. 2008. Underutilization of social insurance among the poor: evidence from the Philippines. *Plos One*. 3(10): e3379.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. Econometrica, 331-341.

Wagstaff, A. 2010. Estimating health insurance impacts under unobserved heterogeneity: the case of Vietnam's health care fund for the poor. *Health Economics*. 19:189-208.

Appendix

Table A1: Robustness to excluding 2006 from sample / treatment period (LFS/QLFS, GHS)

VARIABLES	(1) Health insurance	(2) Medical visit if ill	(3) Health insurance	(4) Medical visit if ill
Public sector	0.221***	-0.184**	0.224***	-0.236**
	(0.00238)	(0.0933)	(0.00276)	(0.113)
Health insurance (predicted)		1.016**		1.214**
		(0.396)		(0.480)
Public sector*2007	0.00958*		0.00368	
	(0.00495)		(0.00517)	
Public sector*2008	0.0502***		0.0443***	
	(0.00379)		(0.00405)	
Sample year	2003-2008		2003-2005 & 2007-2008	
Treatment years	2007-2008		2007-2008	
Observations	329,692	348,466	278,770	294,451
R-squared	0.276		0.278	
F-test	88.81		64.24	
Year effects	Y	Y	Y	Y
Control variables	Y	Y	Y	Y
Bootstrapped s.e.	N	Y	N	Y

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table A2: Robustness to choice of treatment & control groups (LFS/QLFS, GHS)

VARIABLES	(1) Health insurance	(2) Medical visit if ill	(3) Health insurance	(4) Medical visit if ill	(5) Health insurance	(6) Medical visit if ill
Public sector	0.247***	-0.194	0.210***	-0.0799	0.0931***	-0.0242
	(0.00307)	(0.133)	(0.00198)	(0.0596)	(0.00442)	(0.161)
Health insurance (predicted)		1.002*		0.603**		0.610
		(0.540)		(0.266)		(1.748)
Public sector*2006	-0.0275***		-0.00148		-0.0188**	
	(0.00587)		(0.00486)		(0.00884)	
Public sector*2007	0.000584		0.0270***		-0.000217	
	(0.00577)		(0.00476)		(0.00856)	
Public sector*2008	0.0359***		0.0690***		0.00670	
	(0.00455)		(0.00356)		(0.00672)	
Sample years	2003-2008		2000-2008		2003-2008	
Treatment industries	CSP services		CSP services		Transport, storage & communication	
Counterfactual industries	Manufacturing, Construction, Wholesale & retail, Financial services		All others		All others (excl. CSP services)	
Observations	221,853	234,662	472,483	494,467	265,340	279,870
R-squared	0.274		0.270		0.217	
F-test	41.02		132.9		2.578	
Year effects	Y	Y	Y	Y	Y	Y
Control variables	Y	Y	Y	Y	Y	Y
Bootstrapped s.e.	N	Y	N	Y	N	Y

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A3: Effect of predicted health insurance on covariates (LFS/QLFS, GHS)

VARIABLES	(1) Coloured	(2) Indian	(3) White	(4) Male	(5) Age	(7) Years of education	(8) Skilled	(9) Highly skilled
Health insurance (predicted)	0.364	-0.254	-0.236	-0.353	-7.703	-3.542	-0.440	0.0505
	(0.300)	(0.159)	(0.361)	(0.497)	(11.83)	(4.043)	(0.487)	(0.389)
Observations	18,928	18,928	18,928	18,935	18,914	18,804	18,937	18,937
R-squared	0.003	0.001	0.003	0.018	0.003	0.078	0.005	0.131
Year effects	Y	Y	Y	Y	Y	Y	Y	Y

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1