Collaborative Research Project

24 October 2014

Contents

1. Preparation

```
# Set working directory
# setwd('/Users/Meilin/Desktop/Collaborative Social Data')

# For this research, the installation of the following packages is required:
# install.packages('WDI')
# install.packages('betareg')
# install.packages('ggplot2')
# install.packages('magrittr')
# install.packages('repmis')
# install.packages('RJSONIO')
# install.packages('knitr')

# Load all the relevant packages:
library(ggplot2)
library(magrittr)
library(repmis)
library(WDI)
```

```
## Loading required package: RJSONIO
```

```
library(betareg)
library(RJSONIO)
library(knitr)

# For creating our BibTeX libraries we used the following code:
# \bibliography{Literature.bib, Packages.bib}

# The packages used can be automatically cited by following these steps:
# PackagesUsed <- c("WDI", "ggplot2", "magrittr", "repmis", "betareg", "RJSONIO","knitr")
# lapply (PackagesUsed, library, character.only = TRUE)
# knitr::write_bib(PackagesUsed, file = "Packages.bib")
```

2. Introduction

The expiry date of the Millennium Development Goals (MDGs) is just around the corner, meanwhile the post-2015 agenda is being discussed intensively. In this context, it is important to assess the achievement of the MDGs and try to understand why some goals have not been reached.

Reducing HIV prevalence is an important aim of the MDGs. Target 6.A of the MDGs specifies that countries should "have halted by 2015 and begun to reverse the spread of HIV/AIDS" (United Nations (2014)). In most regions of the world this goal has been fulfilled: new HIV infections declined and the overall number of new HIV/AIDS infections per 100 adults (15-49 years old) decreased by 44 per cent between 2001 and 2012 (United Nations (2014)). However, this trend cannot be observed in all 189 member states of the United Nations. On the contrary, HIV/AIDS prevalence has even increased in some countries.

3. Aim, Research Question and Hypotheses

This paper aims to provide evidence to assess why some countries struggle to achieve MDG 6A. We believe that one possible explanation for the failure of some interventions in reducing HIV/AIDS may lie in the lack of a full understanding of the determinants of the disease, which can in turn lead to ill-specified interventions and wrongly targeted campaigns.

The literature reviewed for this paper identifies a myriad of determinants of health. However, there is a gap in the literature when it comes to the determinants of specific diseases. The aim of this paper is to test the applicability of one of the most commonly used theories of determinants of health, to explain the evolution of HIV/AIDS rates. The first hypothesis of this paper is that all general determinants of health (as identified by our reference model) are determinants of HIV/AIDS. By identifying variables that help to explain HIV/AIDS incidence and prevalence, this paper will help move forward the discussion of the determinants of HIV/AIDS .

Furthermore, given that the literature identifies interlinkages between the different determinants of health (see Solar and Irwin (2010)), this paper will evaluate to what extent those linkages are reflected on the micro level. The second hypothesis of this paper is that socioeconomic, environmental and cultural factors can be used to explain individual lifestyle factors.

---

4. Literature Review

Hurrelmann (Hurrelmann (1989), p.76) advocates an interdisciplinary framework for analysing what determines health outcomes. He considers it necessary to use a model that integrates all the aspects of the organism, individual and the environment.

One framework that shows the interaction between individual and environmental factors over time is the salutogenic model developed by Antonovsky in 1979. According to Hurrelmann, Antonovsky's model is a great contribution to interdisciplinary theory, but the downside is its complexity (Hurrelmann (1989)).

A simpler and more common model on the main determinants of health is the "rainbow model", developed by Dahlgren and Whitehead (Dahlgren and Whitehead (1991), p.11). This model gives an overview of the main health determinants, reflecting the relationship between the individual, its environment and different health outcomes. Individuals are at the centre of the model with a set of fixed biological and genetical preconditions. Building upon these, four layers of influence on health can be identified: individual lifestyle factors, social and community networks, living and working conditions and general socio-economic, cultural and environmental conditions.

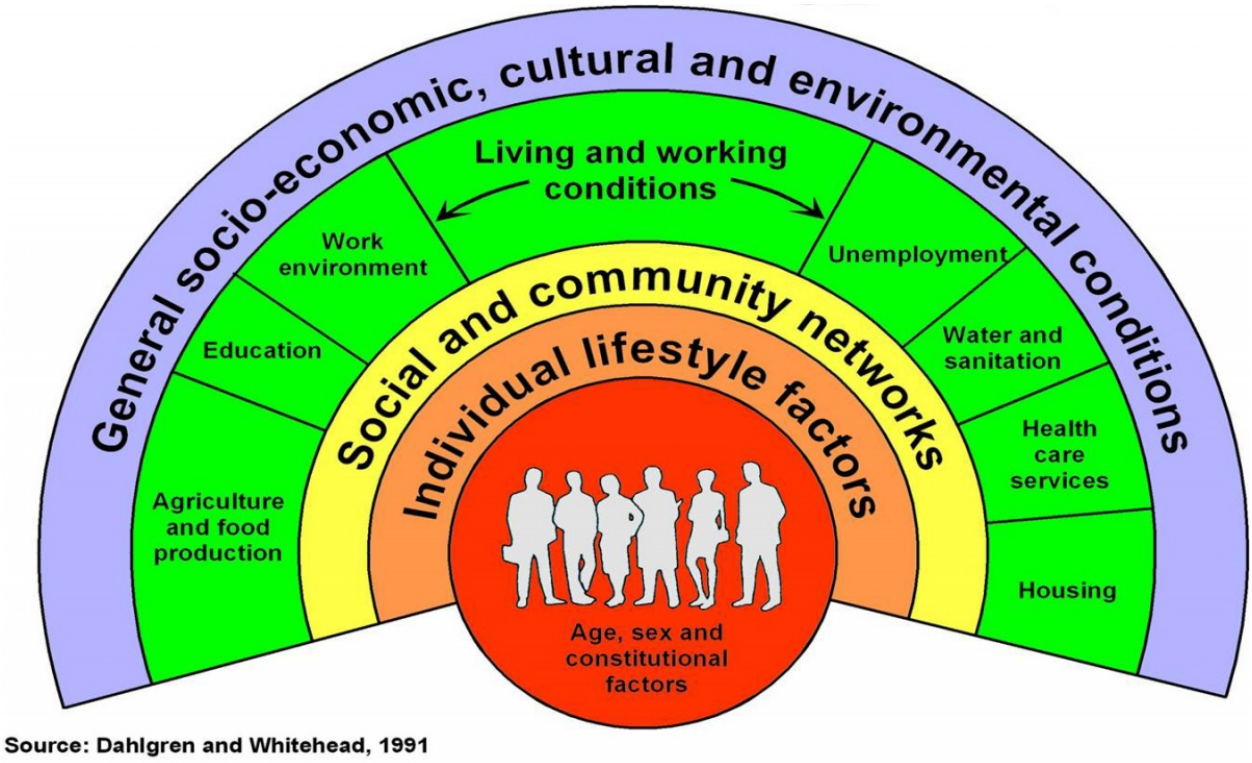

Figure 1: Dahlgren and Whitehead (Dahlgren and Whitehead (1991), p.11). Main health determinants.

---

5. Data Sources

For this research, two main sources of data were explored. Firstly, databases comprising information for several countries were examined in order to increase comparability of the data. Among the most commonly used databases in the field of this research, it is worth mentioning those from UNAIDS, World Bank, Global Fund for AIDS, Tuberculosis and Malaria, WHO, the Institute for Health Metrics and Evaluation, PEPFA and the AIDS Data Hub. Secondly, country-specific data was used to check the trends of some of the most important variables of this research and to enable micro level analysis. These databases include those from the National Bureau of Statistics (from each country under analysis), Demographic Health Surveys (DHS) and the AIDS Indicator Survey (AIS), obtained from the United States Agency for International Development (USAID).

5.1.World Development Indicators (WDI)

The WDI database comprises 1342 indicators clustered in 10 thematic areas that range from health and education to infrastructure and public sector data. Information is available for 214 countries and dates back to 1960. All indicators are available for free at the World Bank website and can be downloaded as an Excel sheet, CSV, tabbed TXT or SDMX. In addition, there is a special R package WDI designed to download and use the data.

WDI have been used in a wide range of fields and HIV/AIDS research is not an exception. For examples of relevant literature that also make use of WDI please see Haacker (2002), Talbott (2007) and Kalemli-Ozcan (2011). A list of all WDI indicators used for this research can be found in the Appendix.

5.2. Demographic Health Surveys (DHS) and AIDS Indicator Survey (AIS)

DHS have a high reputation for collecting accurate and nationally representative household data (World Bank (2014)). Relevant HIV/AIDS literature used the DHS data like Haacker (2002) and World Health

Organization and UNICEF (2014). For the case of Vietnam three datasets for the years 2005, 2002 and 1997 are available in USAID's DHS database. The most recent available DHS dataset will be used for the micro level analysis of this research. The DHS was conducted by the General Statistical Office National Institute for Hygiene and Epidemiology between September and December 2005. There are three core questionnaires in DHS surveys: a household, a women's and a men's questionnaire. The DHS sample for Vietnam counts 6337 households, including 7289 observations for female respondents (women aged 15 to 49) and 6707 observations for male respondents (men aged 15 to 49).

As an extension of the DHS data, the micro analysis also integrates the AIDS Indicator Survey (AIS) for Vietnam, again being conducted in 2005. The AIS provides nationally representative estimates of HIV rates, by collecting blood from representative samples of the population of both men and women in a country. The AIS for Vietnam includes 1675 households. In all households, women and men aged 15-49 are eligible to participate in the survey. The linkage of DHS' HIV test results to the full DHS survey record (without personal identifiers) allows for an in-depth analysis of the socio-demographic and behavioural factors associated with HIV prevalence.

Both datasets can be downloaded as Stata System File, Flat Data, SAS System File and SPSS System File. Individual and AIS datasets can also be downloaded as a hierarchical file format.

It is worth mentioning that special permission was granted in downloading the DHS and AIS databases, representing a potential challenge for the reproducibility of our research.

---

6. Methodology and Data Analysis

In order to answer to our research question, this paper will begin by looking at HIV/AIDS determinants across countries and across time. To do this, a set of indicators operationalizing Dahlgren's determinants of health will be regressed on HIV/AIDS incidence (Model 1) and the statistical significance of each indicator will be assessed. Model 1 consists of four consecutive linear regressions beginning from the outer layer of Dahlgren's model and progressively incorporating the other layers of influential factors on health[1]: Firstly, the impact of socio-economic, cultural and environmental factors on HIV/AIDS incidence will be tested. The second regression will build on this regression and incorporate working and living conditions. The third regression adds a third group of independent variables, namely the social and community network related variables. Finally, the fourth model includes all the above-mentioned variables and incorporates individual lifestyle factors to the right hand side of the regression equation. In all cases, HIV/AIDS prevalence will be controlled for. This regression model will provide statistical information regarding the correlation between the independent variables and HIV/AIDS incidences on the macro level.

Model 2 evaluates the statistically significant indicators of the first model at the micro level. It uses household data from one country that was identified using the visualization tool of the Institute for Health Metrics and Evaluation. This tool displays the countries with the lowest rate of decline in HIV/AIDS incidence and prevalence between 2000 and 2012. For the case study selection, the analysis revealed that Vietnam and South Africa were the worst performers, both in terms of incidence and prevalence (extreme case selection method). Provided the abundance of research on the South African case, this paper works with the Vietnamese case. For the micro analysis, cross-tabulations will be used to assess whether HIV positive individuals score differently than HIV negative ones, on the identified HIV/AIDS determinants that were statistically significant at the macro level.

The last part of this paper (Model 3) will test the second hypothesis by regressing indicators from the inner level of Dahlgren's model against indicators from the outer layer. These examinations will enable us to better understand the channels through which some of the selected indicators impact on HIV/AIDS incidence and prevalence and move closer to the identification of causal mechanisms.

[1]The logic behind the sequencing is to partly control for effects related to the second hypothesis of this paper, i.e. that the outter layers of Dahlgren's model partly explain the innerer levels.

To sum up, this paper will test the following set of equations[2]:

Model 1:

$$1. I_{it} = \beta_0 + \beta_1 SE_{it} + \epsilon_{it}$$
$$2. I_{it} = \beta_0 + \beta_1 SE_{it} + \beta_2 WLC_{it} + \epsilon_{it}$$
$$3. I_{it} = \beta_0 + \beta_1 SE_{it} + \beta_2 WLC_{it} + \beta_3 SCN_{it} + \epsilon_{it}$$
$$4. I_{it} = \beta_0 + \beta_1 SE_{it} + \beta_2 WLC_{it} + \beta_3 SCN_{it} + \beta_4 ILF_{it} + \epsilon_{it}$$

Model 2:

$$P = f(SE, WLC, SCN, ILF)$$

Model 3:

$$ILF_{it} = \beta_0 + \beta_1 SE_{it} + \beta_2 WLC_{it} + \beta_3 SCN_{it} + \mu_{it}$$

For Model 1 we encounter the problem that the dependent variable "HIV/AIDS Incidences" is highly skewed and does not follow a normal distribution. One way to deal with this is to use a log transformation. However, according to Cribari-Neto and Zeileis (Cribari-Neto and Zeileis (2010)) this approach has many disadvantages compared to a beta regression, being an alternative method for dealing with non-normal frequency distributions.

According to the authors, a beta regression model is especially useful, for continuous dependent variables that are restricted to the standard unit interval (0, 1), like HIV/AIDS incidence rates. This paper will therefore use a beta regression (and the corresponding R package "betareg") for the data analysis of the multivariate regression in Model 1. The "betareg" package can be used to perform inference in both fixed and variable dispersion beta regressions.

Given that the number of HIV positive cases in the sample is very limited (only 9 observations), analysing inferential statistics would lead to weak conclusions and highly skewed results. Model 2 thus recurs to cross-tabulations to provide descriptive statistics. Mean values for HIV positive and HIV negative individuals will be compared in order to describe the socio-economic differences between these two groups.

Missing values are expected as the frequency of data collection in all countries covered by the WDI might differ from country to country. If it can be assumed that the variable is stable over time, a missing value in one year will be substituted by the mean of the value before and the value after the missing year. If more than 50% of the values are missing for all the countries, the variable will be dropped.

It is important to mention that some variables will have to be generated by combining existing variables through different mathematical calculations. One example of this is one of the measurements of income distribution that will result from the division of the share of income held by the poorest 10% and the share of income held by the richest 10% of the population.

---

7. Limitations

The above-mentioned methodology faces the following limitations:

- *MULTICOLLINEARITY:* a certain degree of multicollinearity is expected among the indicators used to operationalize the independent variables. Some examples are total school enrolment and female school enrolment, access to water and share of the population living in rural areas and female school enrolment, female literacy and female unemployment. Variation Inflation Factors (VIF) will be calculated to test for multicollinearity and in the case of detection, either a variable will be dropped, outliers removed or all variables will be standardized.

[2]Where I stands for HIV/AIDS incidence, P stands for HIV/AIDS prevalence, SE stands for socioeconomic factors, WLC stands for working and living conditions, SCN stands for social and community networks and ILF stands for individual lifestyle factors.

- *FIXED EFFECTS:* given that religious and other cultural factors have a high incidence on health seeking behaviour and do not change rapidly across time, it is expected to observe fixed effects. In order to deal with these fixed effects, a fixed effects model will be constructed and analysed.
- *HETEROSCEDASTICITY:* most of the indicators identified by this research are ratios, shares of the population or indicators that can take a value from 0 to 1. As a consequence, heteroscedasticity of the data is expected. This paper will deal with heteroscedasticity by working with beta regressions.

8. Appendix

Tables with Macro Model Variables

- General socio-economic, cultural and environmental conditions

Variables	GDP (constant 2005 US$)	GDP per capita, PPP (constant 2011 intern	Poverty gap at $1.25 a day (PPP) (%)
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	2704	2624	658

Variables	Rural population (% of total population)	CO2 emissions (metric tons per capita)	Population growth (annual %)
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	2968	2158	2992

- Living and working conditions

Variables	Access to electricity (% of population)	Health expenditure (total, %GDP)	Health expenditure per capita
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	170	2434	2434

Variables	Hospital beds (per 1,000 people)	Births attended by skilled health staff (	Improved water source (% of population wi
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	1124	785	2533

Variables	Improved sanitation facilities (% of popu	Unemployment, total (% of total labor for	Employment in agriculture (% of total emp
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	2512	2262	1318

Variables	Children in employment, total (% of child	School enrollment, primary (% gross)	Agriculture, value added (% of GDP)
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	202	2082	2331

- Social and community networks

Variables	Unemployment, female (% of female labor f	School enrollment, primary, female (% gro	Literacy rate, adult female (% of females
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	2262	2064	420

Variables	Out-of-pocket health expenditure (% of to	Life expectancy at birth, total (years)	GINI index
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	2434	2609	659

Variables	Income share held by highest 10%	Income share held by lowest 10%	Female headed households(% of households
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)

Variables	Income share held by highest 10%	Income share held by lowest 10%	Female headed households(% of households
Number of countries	214	214	214
Observations	659	659	128

- Individual lifestyle factors

Variables	Condom use, population ages 15-24, female	Condom use, population ages 15-24, male (	Contraceptive prevalence (% of women ages
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	117	101	464

Variables	Immunization, DPT (% of children ages 12-	Immunization, measles (% of children ages	Prevalence of overweight (% of children u
Dataset	World Development Indicators	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214	214
Observations	2655	2655	324

Variables	Smoking prevalence, females (% of adults)	Smoking prevalence, males (% of adults)
Dataset	World Development Indicators	World Development Indicators
Year	1960-2013 (2000-2013)	1960-2013 (2000-2013)
Number of countries	214	214
Observations	135	135

9. Bibliography

Cribari-Neto, Francisco, and Achim Zeileis. 2010. "Beta Regression in R." *J. Stat. Softw.*, 1–24.

Dahlgren, Göran, and Margaret Whitehead. 1991. "Policies and Strategies to Promote Social Equity in Health." *Institute for Futures Studies*.

Haacker, Markus. 2002. "The Economic Consequences of HIV/AIDS in Southern Africa." *IMF Working Paper* 02 (38).

Hurrelmann, Klaus. 1989. "Human Development and Health." *Springer-Verlag*.

Kalemli-Ozcan, Sebnem. 2011. "AIDS, "Reversal" of the Demographic Transition and Economic Development: Evidence from Africa." *Springer-Verlag*.

Solar, O, and A Irwin. 2010. "A Conceptual Framework for Action on the Social Determinants of Health. Social Determinants of Health Discussion Paper." *Policy and Practice*, no. 2.

Talbott, John R. 2007. "Size Matters: The Number of Prostitutes and the Global HIV/AIDS Pandemic." *PLoS ONE*.

United Nations. 2014. "Millennium Development Goals." http://www.un.org/millenniumgoals/aids.shtml.

World Bank. 2014. http://microdata.worldbank.org/index.php/catalog/dhs/about.

World Health Organization, and UNICEF. 2014. "Joint Monitoring Programme for Water Supply and Sanitation." http://www.wssinfo.org/definitions-methods/data-sources/.