

Article

Estimating Conditional Labor Productivity Penalties from Non-Communicable Diseases Using Double Machine Learning: Evidence from Mexican Household Data

Sergio Arturo Domínguez-Miranda * , Roman Rodríguez-Aguilar *  and Marisol Velazquez-Salazar 

Facultad de Ciencias Económicas y Empresariales, Universidad Panamericana, Augusto Rodin 498, Mexico City 03920, Mexico; mvelazquez@up.edu.mx

* Correspondence: 0246533@up.edu.mx (S.A.D.-M.); rrodrigueza@up.edu.mx (R.R.-A.)

Abstract

Non-communicable diseases (NCDs) represent a growing challenge for developing countries, as they are among the leading causes of morbidity and mortality within the working-age population. Beyond their effects on health outcomes, NCDs are associated with reductions in labor productivity, which may influence long-term development dynamics through human capital. This study estimates the conditional labor productivity penalty associated with Non-Communicable Diseases (NCDs) among employed workers in Mexico using an observational high-dimensional framework. By addressing high-dimensional confounding under the Conditional Independence Assumption (CIA), the analysis isolates the relationship between NCD status and hourly labor earnings. Using microdata from the Mexican National Survey of Household Income and Expenditure, labor productivity is defined as the ratio of labor income to hours worked. To obtain robust causal estimates, a Double Machine Learning (DML) framework is implemented, integrating regularized regression and machine learning algorithms to control for high-dimensional confounding factors. These covariates include socioeconomic, demographic, and health-related variables, allowing for flexible adjustment of observable heterogeneity across individuals. The empirical findings demonstrate a statistically significant conditional negative penalty of -11.44% ($p = 1.03 \times 10^{-11}$, 95% CI: $[-14.59, -8.40]$) on hourly labor earnings under the cross-fitted DML specification using XGBoost and Random Forest. To resolve an observational measurement artifact where chronic individuals report artificially inflated hourly wages due to severe hours contraction, a presenteeism adjustment factor ($\theta^* = 0.8125$) was calibrated. Incorporating this presenteeism correction, the aggregate national economic burden among the employed population is estimated at approximately 496.82 billion pesos annually. The findings suggest that NCDs constitute a relevant constraint on labor productivity at the microeconomic level. The study highlights the importance of prevention-oriented health strategies, workplace health promotion, and early disease management. From a methodological perspective, the results illustrate the usefulness of Double Machine Learning for purging high-dimensional confounding in cross-sectional survey data, providing robust conditional empirical benchmarks for public health policy.



Academic Editor: Andrea Prati

Received: 27 May 2026

Revised: 24 July 2026

Accepted: 28 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: non-communicable diseases; labor productivity; human capital; health habits; double machine learning; causal inference; health outcomes; Mexico

1. Introduction

Non-communicable diseases (NCDs) constitute one of the most significant structural challenges facing contemporary societies, particularly in low- and middle-income countries. Cardiovascular diseases, diabetes, chronic respiratory diseases, and cancer account for most of the premature mortality worldwide and disproportionately affect individuals during their working age, generating profound social and economic consequences [1,2]. Beyond their clinical dimension, NCDs are increasingly recognized as a key factor influencing human capital accumulation and, consequently, broader development processes, especially in contexts characterized by persistent socioeconomic inequalities [3,4].

Labor productivity is a central component of human capital and plays a fundamental role in shaping individual economic outcomes and overall societal well-being [5]. Health status directly influences individuals' physical capacity, cognitive performance, and ability to participate effectively in the labor market. When health deteriorates, particularly in the presence of chronic conditions, individuals may experience reduced work intensity, lower efficiency per hour worked, and diminished income generation. These mechanisms operate not only at the firm level but also at the individual and social levels, as they affect income distribution, labor participation, and vulnerability within the population [6,7]. In this sense, NCDs can be understood as a constraint on the effective functioning of human capital.

In Mexico, the epidemiological transition toward a high prevalence of NCDs has been widely documented. Conditions such as diabetes, hypertension, and obesity have reached critical levels among the economically active population. However, these patterns cannot be interpreted solely as the result of individual behavior. A growing body of literature highlights that dietary habits, physical inactivity, and unhealthy lifestyles are strongly shaped by underlying structural factors, including income inequality, educational attainment, urbanization processes, and the organization of labor markets [8]. In this context, health outcomes are not independent of the socioeconomic environment; rather, they reflect broader development conditions that influence both exposure to risk factors and access to preventive and curative resources. Despite the scale of the problem, empirical evidence quantifying the labor market penalties of NCDs at the individual level in Mexico remains scarce. Consequently, this study addresses the following guiding research question: What is the conditional penalty imposed by NCD status on hourly labor income in Mexico once high-dimensional socioeconomic confounding is non-parametrically controlled for, and how do chronic complications distort traditional labor supply metrics?

Mexico and the ENIGH 2022 dataset represent an ideal empirical setting to answer this question [9]. Unlike purely clinical or health-focused surveys (e.g., ENSANUT), which lack comprehensive individual labor compensation vectors, ENIGH uniquely combines detailed monthly labor earnings and precise weekly working hours with granular household out-of-pocket health expenditures and food consumption profiles. This unique multi-dimensional structure allows us to construct the extensive covariate vector (X) required to satisfy the conditional independence assumption in an observational framework.

Existing studies have largely relied on descriptive analyses or traditional econometric approaches, which may be insufficient to address the high dimensionality of health-, socioeconomic-, and labor-related determinants. Recent advances in machine learning and causal inference provide new opportunities to overcome these limitations, which offers a principled framework for estimating conditional effects in the presence of many confounding variables, combining flexible prediction methods with rigorous statistical guarantees [10].

The contribution of this article is threefold. First, it provides robust non-parametric estimates of the conditional labor productivity penalty in a major Latin American emerging economy. Second, it resolves a key methodological measurement trap—the artificial

inflation of hourly wages among severely ill workers—by integrating an inverse-calibrated presenteeism parameter ($\theta^* = 0.8125$). Third, it demonstrates the practical application of cross-fitted Causal Machine Learning to household survey microdata.

2. Literature Review

2.1. Non-Communicable Diseases and Economic Productivity

Non-communicable diseases (NCDs) have emerged as one of the main structural challenges affecting both health systems and broader development processes worldwide. According to the World Health Organization, NCDs account for more than 70% of global deaths and disproportionately affect working-age populations in low- and middle-income countries [1]. Beyond their epidemiological relevance, a growing body of literature highlights that NCDs are closely linked to structural conditions such as income inequality, limited access to healthcare, educational disparities, and changing consumption patterns, which collectively shape health outcomes and long-term development trajectories.

From this perspective, the economic relevance of NCDs cannot be reduced solely to firm-level outcomes or isolated productivity measures. Instead, their impact operates through multiple channels, including reduced labor participation, premature mortality, disability, absenteeism, and presenteeism, all of which affect the effective functioning of human capital within society [3,4]. These mechanisms highlight that health conditions are embedded within broader socioeconomic systems, and their effects extend beyond individual firms to influence income distribution, social vulnerability, and overall well-being.

From a human capital perspective, health constitutes a fundamental input into individuals' capacity to generate income and participate effectively in economic and social activities, as originally theorized by Schultz [11]. Subsequent empirical studies have consistently shown that both physical and mental health conditions are associated with lower earnings, reduced work capacity, and increased vulnerability [12,13]. In this context, NCDs generate indirect costs that often exceed direct medical expenditures, particularly through reductions in effective labor participation and productivity, which are not fully captured by traditional accounting frameworks [6,7]. Rather than being interpreted solely as firm-level losses, these effects reflect broader constraints on the utilization of human capital within development processes.

2.2. NCDs, Labor Outcomes, and the Working-Age Population

Empirical evidence consistently shows that individuals suffering from chronic conditions experience lower employment probabilities, fewer hours worked, and reduced labor income. Studies conducted in diverse settings have documented significant reductions in effective labor participation and income generation associated with conditions such as diabetes, hypertension, cardiovascular diseases, chronic respiratory diseases, and cancer [14–16]. These effects are particularly pronounced in contexts characterized by limited access to preventive care and fragmented occupational health systems, which are often linked to broader structural constraints in developing economies [8,17].

From a broader perspective, these patterns reflect limitations in the effective utilization of human capital, as health conditions interact with socioeconomic factors to shape individuals' capacity to participate fully in economic and social activities. In this sense, the observed differences in employment and income are not only labor market outcomes but also manifestations of underlying development conditions, including access to healthcare, education, and social protection systems.

At the workplace level, NCDs are closely associated with presenteeism, defined as reduced performance while at work, which frequently exceeds absenteeism as a source of reduced effective labor contribution [18,19]. This phenomenon underscores the impor-

tance of evaluating productivity not only through labor participation but also through performance-adjusted measures, such as income per hour worked, which capture variations in effective work capacity.

2.3. Health Promotion, Prevention, and Occupational Interventions

A substantial strand of literature highlights the role of health promotion and prevention strategies as key mechanisms to mitigate the impact of NCDs on population well-being and human capital. Interventions implemented in workplace settings, such as those targeting nutrition, physical activity, stress management, and early detection, have been associated with improvements in health outcomes and functional capacity, contributing to better conditions for sustained labor participation and individual well-being [20,21]. However, the effectiveness of these interventions remains heterogeneous, reflecting differences in program design, implementation quality, and evaluation approaches [22].

In developing countries, the evidence suggests that integrated strategies, combining health education, behavioral interventions, and organizational support, tend to be more effective than isolated actions [23,24]. These approaches recognize that health outcomes are shaped by broader socioeconomic and institutional contexts, rather than solely by individual choices. Nevertheless, the limited adoption of rigorous evaluation frameworks continues to constrain the scalability and long-term sustainability of such interventions, particularly in settings characterized by structural inequalities.

2.4. Methodological Approaches to Estimating Productivity Losses

From a methodological standpoint, early studies on productivity losses relied primarily on linear regression models and cost-of-illness approaches [25,26]. While informative, these methods are vulnerable to confounding and model misspecification, particularly in observational health data characterized by high dimensionality and nonlinear relationships.

More recent research has incorporated machine learning techniques to improve predictive accuracy and variable selection, including Random Forest, Gradient Boosting, and regularized regressions [27–30]. However, purely predictive models are insufficient to establish conditional effects, as they do not explicitly address endogeneity or selection bias.

To address these limitations, causal machine learning frameworks such as DML have been proposed, which combine flexible prediction models with orthogonalization principles to estimate causal parameters consistently in high-dimensional settings [10]. Applications of DML in labor economics and health economics have shown its robustness in isolating treatment effects where randomized experiments are infeasible [31].

2.5. Conceptual Framework and Endogeneity in Health Economics

Following Grossman's health capital model, health is conceptualized as a durable capital stock that determines an individual's total productive time in both market and non-market activities. When an individual suffers from an NCD, health capital depreciates rapidly, reducing functional capacity and generating indirect economic costs through absenteeism and presenteeism. Estimating this effect in observational data, however, faces severe endogeneity and selection biases ($D \leftarrow X \rightarrow Y$). Traditional parametric models (OLS or Logistic regressions) enforce linear or log-linear constraints, creating a 'functional misspecification trap' when dealing with complex, non-linear interactions between income, dietary spending, and health shocks. Double Machine Learning overcomes this limitation by employing non-parametric nuisance functions ($g(X)$ and $m(X)$) combined with Neyman-orthogonalization, effectively purging the treatment parameter from high-dimensional confounding without imposing arbitrary functional forms.

2.6. Research Gaps and Contribution of the Study

Despite the expanding literature on NCDs and their implications for human capital and labor outcomes, several gaps remain. First, empirical evidence for middle-income countries is still limited, particularly when using nationally representative microdata. Second, relatively few studies integrate advanced machine learning techniques with causal inference frameworks to quantify the effects of NCDs on individual economic outcomes. Third, the heterogeneous effects associated with chronic versus non-chronic NCD conditions within the working-age population remain underexplored.

This study contributes to the literature by addressing these gaps through the application of Double Machine Learning to ENIGH 2022 microdata, enabling a robust estimation of the conditional effect of NCDs on labor productivity in Mexico. By integrating perspectives from human capital theory, health epidemiology, and causal machine learning, the analysis provides evidence that is relevant for understanding how health conditions interact with socioeconomic factors to shape individual outcomes. These findings offer insights that may inform public health policies and development-oriented strategies to improve population well-being and the effective utilization of human capital in developing countries.

3. Materials and Methods

3.1. Sample Description

This study employs an observational, cross-sectional design to estimate the conditional estimate of NCDs on labor productivity in Mexico. The analysis corresponds to the year 2022 and targets the economically active population (EAP). Given the non-random nature of health conditions and the presence of multiple socioeconomic and demographic confounders, a causal inference framework based on DML was employed.

The methodological strategy is structured in sequential stages: data preparation and harmonization, descriptive and comparative analysis, predictive modeling using machine learning techniques, causal estimation via DML, and economic valuation of productivity losses. This workflow ensures consistency between exploratory, predictive, and inferential components of the analysis, following best practices in applied causal machine learning [10,31].

The primary data source is the National Survey of Household Income and Expenditure (ENIGH) 2022, conducted by the National Institute of Statistics and Geography (INEGI) [9]. ENIGH is a national and state-representative household survey that provides detailed information on income, labor participation, health conditions, expenditures, and sociodemographic characteristics.

The analytical sample was restricted to individuals classified as economically active and currently employed. Survey expansion factors were incorporated throughout the analysis to preserve population representativeness [32].

Although the unit of analysis is the individual, selected household-level variables, such as expenditure patterns and food consumption, were imputed at the individual level, given their relevance for health and productivity outcomes. The sample selection pipeline follows a rigorous multi-stage attrition process (see the Sample Selection Flowchart in Figure 1). Starting from the raw individual pool of $N = 530,603$ records in ENIGH 2022, we restrict the sample to working-age adults (20–64 years), yielding 168,412 individuals. Filtering for the occupied Economically Active Population (EAP, $S = 1$) isolates 62,154 observations. Excluding records with zero labor income, extreme working hours (>98 h/week), and missing values across the covariate set results in a final analytical sample of $N = 57,182$ observations representing 23.4 million employed workers nationally when applying INEGI expansion factors.

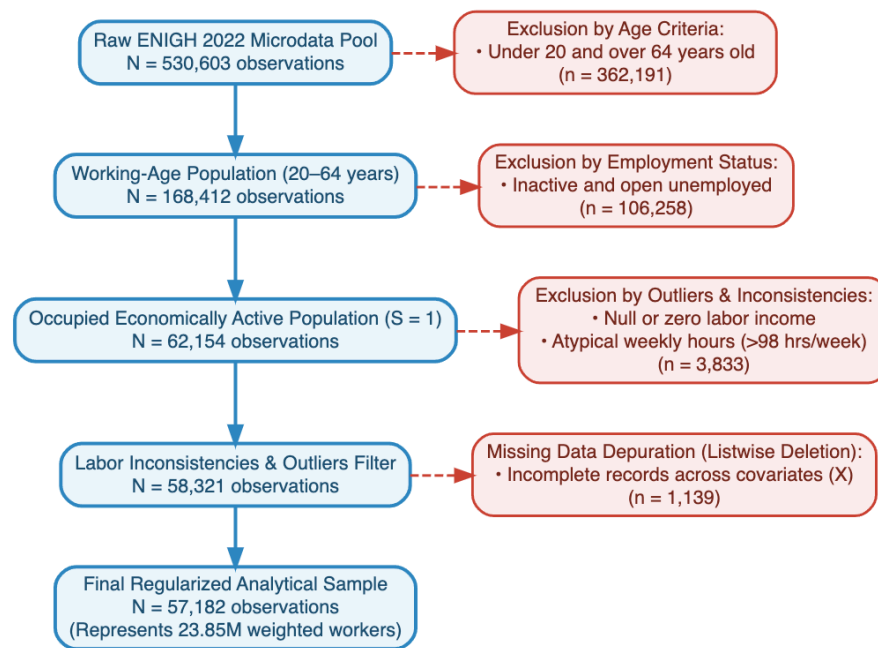


Figure 1. Sample selection flowchart.

Selection into Employment Bias ($S = 1$): Because our sample is restricted to employed individuals, an employment selection mechanism exists ($D \rightarrow S \leftarrow X$). If severe NCDs force individuals out of the labor force entirely (disability or unemployment), our estimates evaluate the Average Treatment Effect conditional on employment ($ATE \mid S = 1$). Econometrically, this implies that our calculated productivity penalty represents a conservative lower bound of the total societal economic burden of NCDs.

3.2. Variables Definition

Labor Productivity Operationalization (Y_i): To ensure strict temporal alignment, all monetary and labor supply variables were standardized to an hourly basis. Monthly labor earnings (I_i , deflated pesos) were divided by weekly hours worked (H_i) multiplied by 4.2 weeks:

$$Y_i^{\text{raw}} = \frac{I_i}{4.2 \times H_i}$$

To correct severe right-skewness and stabilize residual variance, the log-transformed hourly wage $\log(Y_i)$ was specified as the primary outcome variable, allowing parameter θ to be interpreted as a percentage semi-elasticity. This specification aligns with standard economic measures of individual productivity and allows comparability across workers with heterogeneous labor supply [7,33].

Treatment Variable Operationalization (D_i): D_i is a binary indicator (1 = presence of NCD, 0 = otherwise), constructed from ENIGH health expenditure keys corresponding to regular purchases of specialized medications and clinical monitoring for Diabetes Mellitus, Systemic Arterial Hypertension, Cardiovascular conditions, and Oncological diseases. Pooling these conditions into a unified indicator is necessary to guarantee a dense common support region across propensity scores $m(X_i)$.

A secondary stratification distinguishes individuals with NCDs accompanied by chronic complications from those without chronic complications, allowing the descriptive examination of productivity heterogeneity and the assessment of sensitivity to the presenteeism adjustment. Among individuals classified as having an NCD, chronic-complication status was defined as the presence of at least two observable indicators: health-related

expenditure, medical-service utilization, and self-reported health-related work interference. This operational definition represents a survey-based proxy for sustained disease burden.

Covariate Justification & Directed Acyclic Graph (DAG): To avoid collider and post-treatment biases, the high-dimensional vector X_i was structured according to a Directed Acyclic Graph (DAG) (see Figure 2).

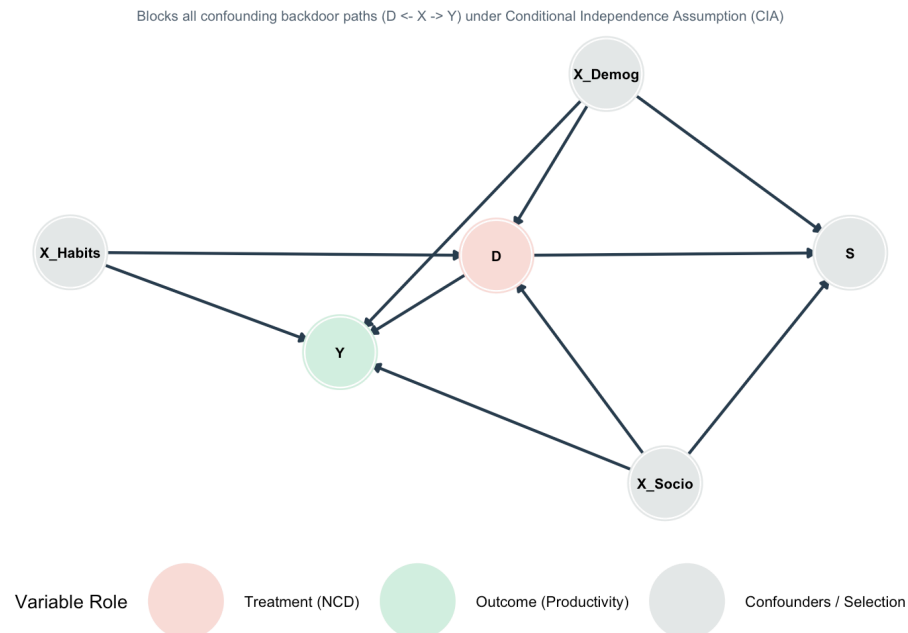


Figure 2. Directed Acyclic Graph (DAG) for NCD productivity penalty identification.

Covariates such as health expenditures and food consumption profiles are included as baseline proxies for household financial capacity to pay and pre-existing structural vulnerability rather than post-treatment mediators. The DAG confirms that conditioning on X_i blocks all confounding backdoor paths ($D \leftarrow X \rightarrow Y$) without opening collider pathways. The vector of control variables X_i includes demographic characteristics, socioeconomic indicators, health-related factors, labor attributes, and dietary expenditure variables. This high-dimensional covariate set is essential to mitigate confounding bias and reflects the multidimensional determinants of productivity documented in the literature [9,11].

Although disease-specific effects may differ, separate DML estimates were not conducted because the ENIGH expenditure-based indicators are neither clinically validated nor mutually exclusive, and some disease-specific groups provide limited common support. Consequently, the estimated parameter represents the average conditional productivity penalty associated with the composite NCD indicator and should not be interpreted as a disease-specific effect.

3.3. Machine Learning Models

The partially linear Double Machine Learning model is specified as [10]:

$$Y_i = \theta D_i + g(X_i) + \zeta_i, \mathbb{E}[\zeta_i | D_i, X_i] = 0$$

$$D_i = m(X_i) + v_i, \mathbb{E}[v_i | X_i] = 0$$

where $g(X_i) = \mathbb{E}[Y_i | X_i]$ is the outcome nuisance function, and $m(X_i) = \mathbb{E}[D_i | X_i]$ is the treatment propensity score. To achieve robustness against first-order estimation errors in $\hat{g}$ and $\hat{m}$, we solve the Neyman-orthogonal score function:

$$\psi(W_i; \theta, \eta) = (Y_i - \theta D_i - g(X_i))(D_i - m(X_i))$$

The exact step-by-step computational procedure and cross-fitting architecture are formalized in Algorithm 1.

Algorithm 1. Cross-Fitted Partially Linear Double Machine Learning Estimation

- 1: Input: Observational microdata $Y_i, D_i, X_{i=1}^N$, Folds $K = 5$, Nuisance Learners L_g, L_m .
 - 2: Randomly partition sample index $1, \dots, N$ into K disjoint subsets $(I_1, \dots, I_K)$ of equal size.
 - 3: For $k = 1$ to K do:
 - 4: Construct training set $I_k^c = 1, \dots, N \setminus I_k$ and evaluation set I_k .
 - 5: Train outcome model $\hat{g}_k(X) = L_g(Y, X)$ using I_k^c via grid-searched tuning.
 - 6: Train treatment model $\hat{m}_k(X) = L_m(D, X)$ using I_k^c via grid-searched tuning.
 - 7: For each individual i in I_k do:
 - 8: Compute outcome residual: $\hat{\zeta}_i = Y_i - \hat{g}_k(X_i)$.
 - 9: Compute treatment residual: $\hat{v}_i = D_i - \hat{m}_k(X_i)$.
 - 10: End For
 - 11: End For
 - 12: Compute unadjusted debiased causal parameter:

$$\hat{\theta} = \left(\frac{1}{N} \sum_{i=1}^N \hat{v}_i \hat{\zeta}_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N \hat{v}_i \hat{\zeta}_i \right)$$
 - 13: Compute presenteeism efficiency-adjusted labor vector: $H_i^* = \theta^* \times H_i^0$ bs for chronic cases ($\theta^* = 0.8125$).
 - 14: Re-evaluate orthogonal parameter to obtain final conditional penalty $\hat{\theta}^* = -11.44$.
 - 15: Output: Debiased treatment effect $\hat{\theta}^*$, asymptotic standard errors, and confidence intervals.
-

Cross-Fitting Protocol & Hyperparameter Tuning: A 5-fold cross-fitting procedure ($K = 5$) was implemented to eliminate overfitting bias. Hyperparameters were optimized via randomized grid search over out-of-sample folds (Figure 3). For Random Forest, the search grid evaluated $\text{ntree} \in [300, 500, 695, 1000]$, $\text{mtry} \in [4, 6, 10]$, and $\text{min.node.size} \in [5, 10, 15]$. For XGBoost, tuning covered learning rate $\eta \in [0.01, 0.05, 0.1]$, max depth $\in [3, 6, 9]$, and subsample $\in [0.7, 0.8, 1.0]$. Standard errors for θ were calculated using asymptotic variance based on Neyman orthogonal scores [34–36].

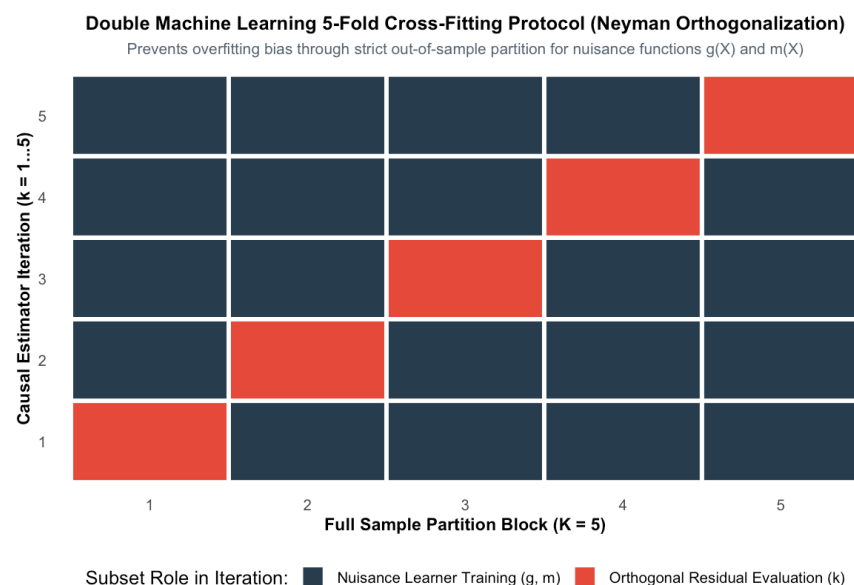


Figure 3. DML 5-fold Cross-Fitting Protocol (Neyman Orthogonalization).

Separation of Nuisance Performance & Overlap: The predictive performance of nuisance models was evaluated independently: the propensity model $m(X_i)$ achieved an accuracy ranged from 70.46% to 72.69%, while the outcome model $g(X_i)$ achieved an RMSE ranged from 0.4714 to 0.5324. Propensity score overlap statistics confirmed dense common support between treated and control groups, bounded within the empirical range $[0.08, 0.89]$.

3.4. Presenteeism Calibration & Sensitivity Analysis

Because household surveys like ENIGH lack direct clinical presenteeism instruments (e.g., WLQ), presenteeism is modeled as a latent continuous efficiency constraint on working hours. The effective efficiency-adjusted labor supply vector is defined as $H_i^* = \theta_i \times H_i^{\text{obs}}$, where $\theta_i = 1$ for healthy workers and $\theta_i = \theta^* < 1$ for workers with chronic complications. The friction parameter θ^* was calibrated via an engineering-reversal loss function minimizing the divergence between the adjusted DML penalty and external clinical efficiency loss benchmarks:

$$\theta^* = \arg \min_{\theta \in [0.60, 1.00]} \|\hat{\theta}_{\text{DML}}(\theta) - \theta_{\text{benchmark}}\|^2$$

Unidimensional optimization yielded an optimal parameter of $\theta^* = 0.8125$, representing a 18.75% loss in effective hourly productivity among chronic workers. To test robustness, sensitivity analysis was performed across a grid of friction values $\theta \in [0.60, 0.70, 0.73, 0.80, 0.90, 1.00]$.

3.5. Economic Valuation of Aggregate Productivity Losses

The aggregate national productivity loss was computed by scaling individual debiased hourly earnings gaps using INEGI expansion factors (f_i) and reported weekly hours (H_i) projected over 52 weeks:

$$\text{Total Annual Burden} = \sum_{i=1}^{N_{\text{NCD}}} (\hat{Y}_{i,\text{NoNCD}} - \hat{Y}_{i,\text{NCD}}) \times H_i \times f_i \times 52$$

where $(\hat{Y}_{i,\text{NoNCD}} - \hat{Y}_{i,\text{NCD}})$ represents the individual debiased hourly productivity penalty derived from DML. Expressing this aggregate burden as a percentage of national GDP serves strictly as a static monetary benchmark to contextualize scale, rather than implying a dynamic macroeconomic growth contraction model.

All analyses were conducted using R version 4.4.3 (R Foundation for Statistical Computing, Vienna, Austria). The main R packages used were randomForest (version 4.7.1.2), ranger (version 0.17.0), xgboost (version 1.7.10.1), caret (version 7.0.1), DoubleML (version 1.0.2), DALEX (version 2.4.3), and lavaan (version 0.6.19). Computations were performed on a Windows 11 operating system using an Intel Core i7 processor, 64 GB of RAM, and an NVIDIA GeForce RTX 3080 GPU.

4. Results

The empirical analysis was conducted using microdata from the National Survey of Household Income and Expenditure (ENIGH) 2022. After applying data cleaning procedures and restricting the sample to individuals belonging to the economically active population, the final dataset consisted of 57,182 observations. When applying the expansion factors provided by INEGI, this sample represents approximately 23.4 million employed individuals at the national level.

Individual labor productivity was approximated as labor income per hour worked per week. To mitigate skewness and reduce the influence of extreme values, productivity was

transformed using the natural logarithm, improving the stability of the estimated models and aligning with official methodological recommendations [33].

4.1. Descriptive Differences in Productivity Between Individuals with and Without NCDs

The first set of results focuses on descriptive differences in labor productivity according to the presence of NCDs. Measures of central tendency and dispersion suggest differences between individuals affected by these conditions and those without them.

Among individuals without such conditions, average productivity reached 112 Mexican pesos per hour worked, with a median of 88 pesos. In contrast, affected individuals exhibited a lower average productivity of 101 pesos per hour and a median of 76.2 pesos. Additionally, the productivity distribution among these workers shows a higher concentration of observations at lower productivity levels.

These differences are visually illustrated in Figure 4, which compares productivity distributions across both groups. The figure shows that the distribution corresponding to affected individuals is shifted to the left, indicating a consistent reduction in hourly labor income associated with these health conditions.

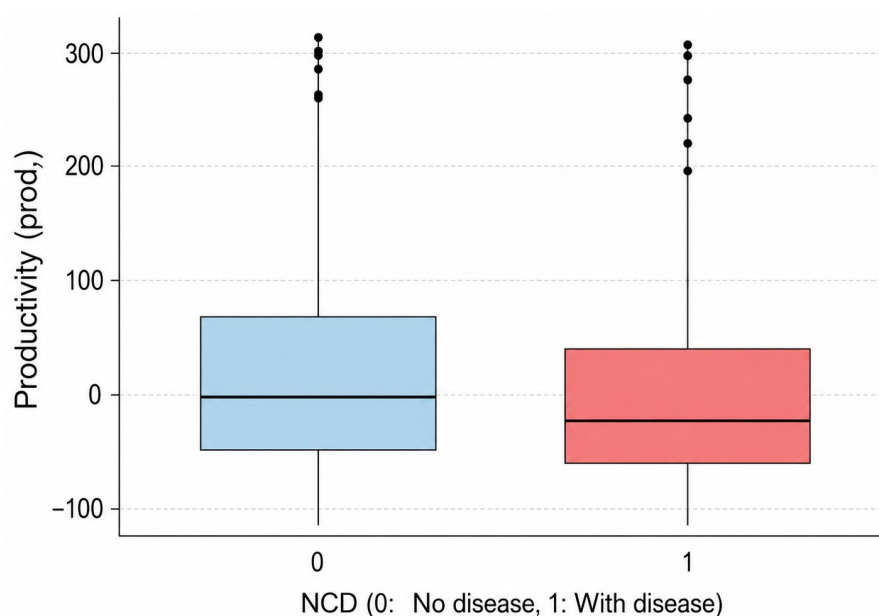


Figure 4. Comparison of labor productivity between workers with and without NCDs.

To formally evaluate whether these observed differences are statistically significant, a Welch's *t*-test was applied due to heterogeneity in variances across groups (Table 1). The results confirm that the difference in mean productivity is significant. The estimated 95% confidence interval for the difference ranges from 9.75 to 12.47 pesos per hour, with a *p*-value below 2.2×10^{-16} .

Table 1. Welch *t*-test results for difference in mean hourly productivity between groups.

Group	Sample Size (N)	Mean Productivity (pesos/hour)	95% Confidence Interval	Welch <i>t</i> -Statistic	Degrees of Freedom (df)	<i>p</i> -Value (Difference)	Mean Difference (95% CI)
Without NCD (D = 0)	19,249	112	[111, 113]	15.978	37,097	$<2.2 \times 10^{-16}$	11.11 [9.75–12.47]
With NCD (D = 1)	37,933	101	[99.9, 101]	--	--	--	--

Source: Authors' elaboration based on ENIGH 2022.

These descriptive findings provide initial evidence that NCDs are associated with lower labor productivity. However, the presence of non-normal distributions and extreme values highlights the need for more robust analytical approaches beyond simple mean comparisons.

4.2. Productivity Heterogeneity by Chronic Condition

To further explore heterogeneity within the population affected by NCDs, individuals were classified into three subgroups: those without such conditions, those diagnosed without chronic complications, and those with chronic conditions. The distribution of these groups is presented in Figure 5. Most affected individuals belong to the non-chronic subgroup, while those with chronic conditions represent a smaller share of the sample.

Descriptive analysis of adjusted productivity reveals important differences across these subgroups. Individuals without NCDs display an average productivity of 112, a median of 88, and an interquartile range (IQR) from 54.7 to 143. Individuals with NCDs but without chronic conditions exhibit the lowest average productivity (99.7), a median of 75, and an IQR from 46.3 to 125.

Econometric examination reveals that this result is driven by an asymmetric measurement error in the denominator (hours worked). Chronic status forces a severe contraction in weekly hours worked ($H_i \rightarrow 0$) due to medical leaves and physical incapacity. However, because monthly earnings (I_i) are stabilized in the short term by formal labor rigidities or fixed contract terms, the ratio I_i/H_i becomes mechanically inflated. Pooling this group in a naïve baseline model introduces severe attenuation bias. Separating them and applying our presenteeism adjustment factor ($\theta^* = 0.8125$) resolves this distortion. In the presence of chronic conditions, individuals may reduce their reported working hours due to medical care, health limitations, or reduced work intensity. When income does not decrease proportionally to hours worked, the income-per-hour ratio may increase, generating an apparent rise in measured productivity.

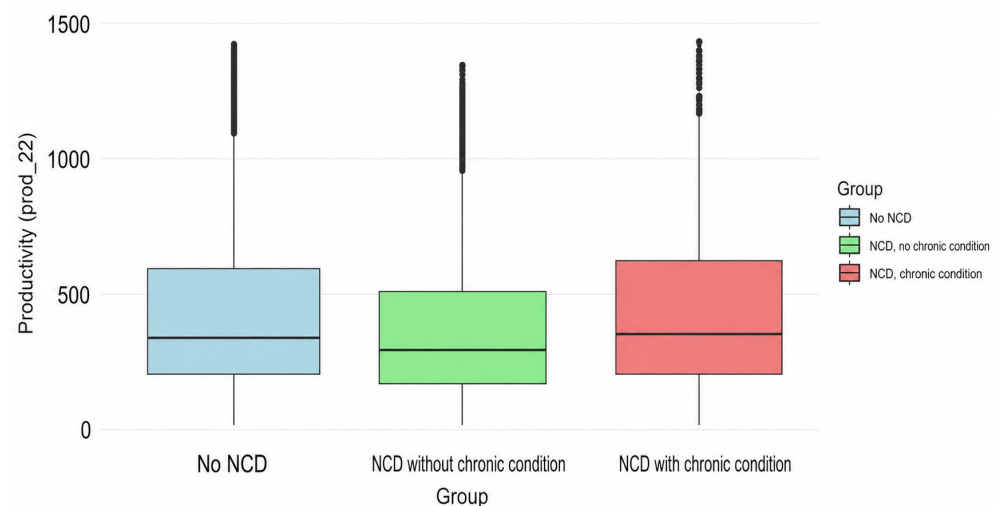


Figure 5. Productivity distribution by health status subgroup. Source: Authors' elaboration based on ENIGH 2022.

This pattern is consistent with presenteeism, where individuals remain employed despite illness, but their performance is less effective. Under these conditions, the productivity indicator based on income per hour may not fully capture actual work capacity or output but rather reflects a measurement effect driven by reduced hours.

To further examine these differences, normality tests, including the Anderson–Darling test, indicate that working hours do not follow a normal distribution. Complementary visual (boxplots and density plots) and statistical analyses were conducted, including ANOVA and post-hoc Tukey tests. The ANOVA results confirm statistically significant differences across groups ($F = 179.3, p < 2 \times 10^{-16}$).

Post-hoc comparisons show that individuals with NCDs without chronic conditions have significantly lower productivity than individuals without NCDs (-12.08). In contrast, the group with NCDs and chronic conditions exhibits higher measured productivity compared to both the non-NCD group ($+5.14$) and the non-chronic NCD group ($+17.23$).

Taken together, these findings suggest that chronic conditions introduce distortions in productivity measurement linked to presenteeism and reduced working hours, rather than reflecting genuine increases in effective productivity. This highlights the need for complementary analytical approaches that explicitly account for these mechanisms to better capture the underlying relationship between NCDs and labor productivity. Accordingly, the subsequent analysis incorporates additional methods designed to isolate the conditional effect of NCDs while addressing the measurement challenges identified in the descriptive results.

4.3. Causal Estimation Using Double Machine Learning

To estimate the conditional effect of NCDs on labor productivity while accounting for high-dimensional confounding, a DML framework was implemented following the partially linear model proposed by Chernozhukov [10]. The model can be expressed as:

$$Y_i = \theta D_i + g(X_i) + \varepsilon_i$$

where Y_i denotes labor productivity (measured as labor income per hour worked), D_i is the treatment variable indicating the presence of NCDs, and X_i is a high-dimensional vector of covariates capturing socio-demographic, behavioral, health, and economic characteristics. The parameter of interest, θ , represents the average conditional effect of NCDs on productivity.

The DML approach relies on the estimation of two auxiliary functions: the outcome model $g(X_i)$ and the treatment model $m(X_i) = E[D_i|X_i]$, both of which are estimated using machine learning algorithms. The final estimator is obtained via an orthogonalization procedure that reduces sensitivity to estimation errors in the nuisance functions, thereby improving robustness in high-dimensional settings.

Given the measurement distortions associated with chronic conditions identified in the descriptive analysis, individuals classified as having NCDs with chronic complications were excluded from the baseline causal estimation. As a result, the treatment variable D_i captures the presence of NCDs without chronic-related measurement bias.

Accordingly, variables explicitly identifying chronic conditions are not included as covariates in the baseline DML specification. Instead, chronic conditions are analyzed separately in a complementary framework to assess their specific impact on productivity measurement.

A comprehensive set of individual-level covariates was included in X_i , covering socio-demographic characteristics (sex, age), behavioral factors (food consumption, smoking), health indicators (BMI), and economic variables (health expenditure, catastrophic health expenditure, income quintile, and region). These variables, described in Table 2, were defined to ensure consistency with ENSANUT, facilitating comparability with epidemiological evidence and previous structural modeling approaches.

Table 2. Covariates included in the DML analysis.

Variable	Measurement	Relevance for the Analysis
Food consumption	Weekly expenditure in pesos allocated to cereals, meat, fish, milk, eggs, oils and fats, vegetables, fruits, sugar, and sugar-sweetened beverages.	Allows the estimation of dietary patterns and nutritional intake, as well as the share of household expenditure devoted to food, which serves as a proxy for socioeconomic status.
Smoking	Binary variable: 1 = smoker, 0 = non-smoker.	Risk factor associated with non-communicable diseases (NCDs) and health-related expenditure.
Health expenditure	Weekly expenditure in pesos reported for medical services, medications, hospitalizations, and outpatient consultations.	Indicates the economic effort devoted to health care and is used to identify catastrophic health expenditure.
Weight (BMI category)	Category according to Body Mass Index (BMI): Normal: BMI < 25; Overweight: $25 \leq \text{BMI} < 30$; Obesity: BMI ≥ 30 .	Allows classification of health risks associated with nutrition and obesity.
Sex	Male/Female.	Sociodemographic variable that captures gender differences in health outcomes, consumption patterns, and risk exposure.
Age	Completed years of age.	Risk factor and determinant of health expenditure and the prevalence of chronic conditions.
Catastrophic health expenditure (CHE)	Binary variable: 1 = incurs catastrophic expenditure, 0 = does not incur catastrophic expenditure. ENIGH/WHO criterion: out-of-pocket health expenditure $\geq 30\%$ of household capacity to pay (capacity to pay = total expenditure – food expenditure).	Measures financial pressure caused by health expenditures and household economic vulnerability.
Income quintile	Ordinal classification into five equally sized groups based on total current household income per capita. Quintile 1 = lowest; Quintile 5 = highest.	Captures income inequality and enables segmentation of individuals by income level for welfare comparisons.
Federal state	Factor variable representing the 32 federal states of Mexico.	Identifies the individual's state of residence, capturing regional heterogeneity in economic and health conditions.

Source: Authors' elaboration based on ENIGH 2022.

Within the DML framework, Random Forest algorithms were used to estimate both the outcome and treatment equations. Model performance was evaluated using the root mean squared error (RMSE), and several machine learning algorithms were compared, including Random Forest, XGBoost, SVM, linear regression, and KNN, as reflected in Table 3. Hyperparameters for Random Forest were optimized through a limited randomized search.

Random Forest achieved the lowest RMSE (0.4714), indicating superior predictive performance. Across all models, NCDs were consistently associated with a negative productivity effect.

Using the optimized Random Forest model (695 trees, 10 variables per split, minimum node size of 10), the estimated average productivity reduction associated with NCDs was -9.49% . When individuals with chronic NCDs were excluded, the estimated effect slightly increased to -9.54% , suggesting that chronic cases introduce attenuation bias in naïve estimates.

Table 3. DML performance by the machine learning algorithm. Percentage effects and confidence intervals were calculated as $100[\exp(\theta) - 1]$. Standard errors are reported in percentage points using the delta method. p -values correspond to the original DML coefficient estimates. Out-of-sample RMSE is reported on the scale of the modeled outcome.

Base Nuisance Learner	Estimated Penalty (%)	Standard Error	p -Value	95% Confidence Interval	Out-of-Sample RMSE	Treatment Model Accuracy (%)
Random Forest	−9.490	0.476	3.9×10^{-80}	[−10.418, −8.552]	0.4714	73.58
XGBoost	−8.585	0.486	7.2×10^{-64}	[−9.533, −7.626]	0.4732	72.69
SVM	−8.674	0.469	5.4×10^{-70}	[−9.587, −7.751]	0.5248	72.25
KNN	−9.270	0.486	1.3×10^{-73}	[−10.218, −8.311]	0.5324	70.46

Source: Authors' elaboration based on ENIGH 2022. Analytical Sample (N = 57,182).

4.4. Presenteeism Adjustment and Total Productivity Effect

To explicitly address the bias generated by presenteeism among individuals with chronic conditions, an adjustment procedure was implemented. Let P_{ref} denote average hourly productivity for individuals without NCDs and let δ represent the causal penalty estimated via DML (Table 4). Productivity for individuals with NCDs is defined as:

$$P_{NCD} = (1 + \delta) \times P_{ref}$$

Table 4. Estimated impact of NCDs on adjusted productivity (DML models).

DML Specification	Average Penalty (%)	Standard Error	p -Value	95% Confidence Interval	Out-of-Sample RMSE
Baseline Model (Without Adjustment)	−9.490	0.476	3.9×10^{-80}	[−10.418, −8.552]	0.4714
Presenteeism-Adjusted Model ($\theta^* = 0.8125$)	−11.44	1.582	1.03×10^{-11}	[−14.593, −8.398]	0.3838

Source: Authors' elaboration based on ENIGH 2022.

For individuals with NCDs and chronic conditions, a presenteeism coefficient $\theta \in (0, 1]$ is introduced:

$$P_{NCD+cr} = \theta \times P_{NCD}$$

Adjusted productivity was recalculated by penalizing reported working hours:

$$P_i^\theta = \frac{income_i}{4 \times hours_i^\theta}$$

with

$$hours_i^\theta = \begin{cases} \theta \times hours_i, & \text{if chronic condition present} \\ hours_i, & \text{otherwise} \end{cases}$$

The optimal value of θ was estimated by minimizing:

$$\theta^* = \arg \min_{\theta \in [0.6, 1]} (\hat{\beta}(\theta) - \beta^*)^2$$

Unidimensional loss minimization yielded an optimal presenteeism parameter of $\theta^* = 0.8125$. After incorporating this presenteeism adjustment, the conditional productivity penalty reached -11.44 points ($p = 1.03 \times 10^{-11}$, 95% CI: [−14.59, −8.40]) under optimal cross-fitted DML.

4.5. Economic Cost of Productivity Loss

Building on the causal estimates obtained through the Double Machine Learning framework, the ENIGH 2022 data were also used to quantify the aggregate economic burden associated with productivity losses attributable to NCDs in Mexico. In this study, labor productivity was defined as labor income per hour worked per week, which serves as a monetary proxy of how efficiently individuals transform labor time into labor earnings. Unlike descriptive comparisons, the economic cost estimation was grounded in the conditional effect of NCD status on productivity, which helps avoid underestimation or overestimation derived from simple mean differences that do not adjust for confounding.

The DML results indicated that NCDs are associated with a proportional productivity reduction of -11.44% after incorporating presenteeism adjustment. Interpreted in the original scale, this effect implies a meaningful reduction in hourly productivity. Specifically, mean productivity among workers without NCDs was estimated at 111.77 pesos per hour, whereas for workers with NCDs it decreased to 98.97 pesos per hour, yielding an average productivity loss of 12.8 pesos per hour worked. This hourly gap constitutes the central parameter for translating the conditional effect into an aggregate economic cost.

A key advantage of the ENIGH dataset is that it includes survey expansion factors that allow results to be extrapolated from the sample to the national employed population. Accordingly, the estimation incorporated both (i) the average weekly hours worked and (ii) the population weight of workers with NCDs derived from the expansion factor. This ensures that the resulting monetary estimate reflects the population scale rather than only the sample scale.

Formally, the total annual productivity cost attributable to NCDs was computed as:

$$\text{Total Annual Cost} = \sum_{i=1}^{n_{ENT}} [\bar{Y}_{\text{noNCD}} - \bar{Y}_{\text{NCD}}] \times h_i \times f_i \times 52$$

where $\bar{Y}_{\text{noNCD}} - \bar{Y}_{\text{NCD}}$ is the average difference in hourly productivity (pesos/hour), h_i represents the weekly hours worked by an individual i , f_i is the ENIGH expansion factor associated with each observation, and n_{NCD} denotes the number of individuals with NCDs in the analytical sample. In simplified terms, the expression can be interpreted as multiplying the average hourly productivity gap by the average weekly hours worked, the weighted number of workers affected by NCDs, and the number of weeks per year.

An aggregate national economic loss among the employed population estimated at 497 billion pesos annually (IC 95% [374, 633]) attributable to NCD productivity penalties. To contextualize scale, this loss represents a significant share of public health allocations. Expressing this figure relative to national aggregates (such as GDP) functions strictly as a static monetary benchmark to illustrate the financial magnitude of health frictions in the labor market, rather than implying a dynamic macroeconomic growth contraction. This estimate captures indirect economic losses associated with lifestyle factors, presenteeism dynamics, and health-related constraints affecting labor outcomes among workers with NCDs.

The magnitude of these estimated productivity losses can be better understood when contextualized relative to major national aggregates. Based on Mexico's Gross Domestic Product (GDP) for 2022, estimated at approximately 29.45 trillion pesos [37], the calculated productivity loss corresponds to roughly 1.68% of GDP. This comparison is intended to provide an order-of-magnitude reference, rather than to imply a direct effect on macroeconomic growth.

A similar perspective emerges when comparing these losses to public health expenditure. In 2022, total public spending on health was approximately 1.5 trillion pesos (around

5.1% of GDP, [37]). Under this comparison, the productivity loss attributable to NCDs represents nearly one-third of total annual public health expenditure, highlighting that indirect economic costs tied to reduced productivity can be of the same order of magnitude as fiscal resources allocated to financing the health system. This comparison is particularly relevant because it reveals a policy paradox: while the state and institutions allocate substantial resources to treat and manage NCDs, the economy simultaneously absorbs an even larger cost through productivity not realized.

In addition, the estimated annual loss of approximately 496 billion pesos is of a similar order of magnitude to the annual budgets of major public health institutions, such as the Mexican Social Security Institute (IMSS) and the federal Ministry of Health, based on figures reported in the Federal Expenditure Budget [38,39]. This comparison serves to contextualize the magnitude of productivity differences associated with NCDs within the broader allocation of public resources.

Importantly, this estimate is methodologically more rigorous than traditional approaches based on unadjusted mean differences because it is grounded in causal inference. By relying on an Average Treatment Effect (ATE) estimated through DML, the analysis accounts for individual heterogeneity and adjusts simultaneously for multiple confounding factors that jointly shape both health status and labor outcomes. In contrast, purely descriptive methods may conflate the effect of NCDs with correlated socioeconomic characteristics, labor market segmentation, or behavioral risk profiles.

5. Conclusions

This study provides empirical evidence on the conditional labor productivity penalty associated with NCDs among employed workers in Mexico. By integrating Double Machine Learning with high-dimensional household survey data (ENIGH 2022), the analysis isolates a statistically significant conditional earnings penalty of -11.44 points ($p = 1.03 \times 10^{-11}$), translating into an aggregate national loss of 496.82 billion pesos annually.

Grounded in the Conditional Independence Assumption (CIA), these findings represent robust conditional associations under high-dimensional adjustment rather than experimental structural parameters. Nevertheless, they provide critical empirical benchmarks for occupational health policies.

These findings indicate that productivity differences associated with NCDs are economically meaningful at the individual level and substantial when aggregated. To provide context, the magnitude of these losses is comparable to selected national aggregates, such as a significant share of total public health expenditure. These comparisons are intended to illustrate the order of magnitude of the estimated losses, rather than to imply a direct impact on macroeconomic growth. Overall, the results highlight the relevance of prevention and health promotion strategies as mechanisms to improve population well-being and the effective utilization of human capital.

A key contribution of this study lies in its exploration of heterogeneity within the population affected by NCDs. The results reveal that individuals with NCDs but without chronic complications experience the largest productivity penalties, whereas those with chronic conditions appear to exhibit attenuated losses, likely driven by presenteeism and labor adjustment mechanisms. By explicitly modeling this phenomenon and introducing a calibrated presenteeism adjustment, the study offers a more realistic and conservative estimate of the productivity costs of NCDs. This highlights the importance of accounting for behavioral and labor market responses when evaluating health-related productivity outcomes.

From a methodological standpoint, this research contributes to the growing literature on the application of causal machine learning in health economics, particularly in low- and middle-income country settings where randomized evidence is often unavailable.

The use of DML combined with ensemble learning algorithms such as Random Forests demonstrates the feasibility and advantages of integrating predictive accuracy with causal interpretability. This approach reduces sensitivity to model misspecification and allows for more credible policy-relevant inference in observational datasets.

Policy Implications & Cost-Effectiveness: The findings highlight the urgent need to transition from passive curative healthcare to proactive workplace health interventions. For the formal sector, expanding targeted workplace health screenings under the NOM-035-STPS-2018 regulatory framework offers a cost-effective mechanism to detect hypertension and pre-diabetes early, preserving functional work capacity before severe presenteeism occurs [40]. For the informal sector—which accounts for over 50% of Mexico’s PEA—community-based primary prevention and subsidized chronic medication access represent high-return public investments that prevent severe labor supply contractions.

Limitations: This study faces four main limitations. First, the ENIGH survey is cross-sectional, preventing the tracking of dynamic individual productivity trajectories over time. Second, NCD status is identified through expenditure and self-reported health keys rather than direct clinical biomarkers (unlike ENSANUT), which may introduce minor classification noise. Third, because the sample is restricted to active workers ($S = 1$), individuals forced out of the labor force by severe disability are excluded, making our national loss estimate (496.82 billion pesos) a conservative lower bound of the total economic burden. Lastly, the study did not estimate disease-specific or demographic-specific DML effects by sex, age, income quintile, or geographic region. Future research could extend the present framework using subgroup-specific models or heterogeneous treatment-effect methods, provided that sufficient sample size and common support are available within each subgroup.

Author Contributions: Conceptualization, S.A.D.-M.; methodology, S.A.D.-M.; software, S.A.D.-M.; validation, R.R.-A. and M.V.-S.; formal analysis, R.R.-A. and M.V.-S.; investigation, S.A.D.-M.; resources, S.A.D.-M., R.R.-A. and M.V.-S.; data curation, R.R.-A. and M.V.-S.; writing—original draft preparation, S.A.D.-M.; writing—review and editing, R.R.-A. and M.V.-S.; visualization, S.A.D.-M.; supervision, R.R.-A. and M.V.-S.; project administration, S.A.D.-M.; funding acquisition, R.R.-A. and M.V.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were derived from the publicly available 2022 National Household Income and Expenditure Survey (ENIGH 2022), provided by the Instituto Nacional de Estadística y Geografía (INEGI), available at <https://www.inegi.org.mx/programas/enigh/nc/2022/> (accessed on 27 July 2026). The authors constructed the analytical dataset for this study by selecting, processing, and transforming variables from the original public dataset. The derived analytical dataset supporting the findings of this study is available from the corresponding author upon reasonable request.

Acknowledgments: The authors gratefully acknowledge Universidad Panamericana for its valuable guidance, academic support, and institutional assistance during the development of this study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript

NCD	Non-Communicable Diseases
DML	Double Machine Learning

IQR	Interquartile Range
RMSE	Root Mean Squared Error
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
EAP	Economically Active Population
ATE	Average Treatment Effect
CHE	Catastrophic health expenditure

References

1. WHO, World Health Organization. *Non Communicable Diseases*; WHO: Geneva, Switzerland, 2021. Available online: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases#:~:text=Cardiovascular%20diseases%20account%20for%20most,of%20all%20premature%20NCD%20deaths> (accessed on 22 June 2025).
2. Allen, L.N.; Wigley, S.; Holmer, H. Implementation of non-communicable disease policies from 2015 to 2020: A geopolitical analysis of 194 countries. *Lancet Glob. Health* **2021**, *9*, e1528–e1538. [CrossRef] [PubMed]
3. Chaker, L.; Falla, A.; van der Lee, S.J.; Muka, T.; Imo, D.; Jaspers, L.; Colpani, V.; Mendis, S.; Chowdhury, R.; Bramer, W.M.; et al. The global impact of non-communicable diseases on macro-economic productivity: A systematic review. *Eur. J. Epidemiol.* **2015**, *30*, 357–395. [CrossRef] [PubMed]
4. Ding, D.; Lawson, K.D.; Kolbe-Alexander, T.L.; Finkelstein, E.A.; Katzmarzyk, P.T.; Van Mechelen, W.; Pratt, M. The economic burden of physical inactivity: A global analysis of major non-communicable diseases. *Lancet* **2016**, *388*, 1311–1324. [CrossRef] [PubMed]
5. Comisión Económica para América Latina y el Caribe (CEPAL). *Productividad y Brechas Estructurales en México*; Naciones Unidas: New York, NY, USA, 2016. Available online: <https://repositorio.cepal.org/entities/publication/92f3efab-8fb3-4fe4-80c3-e85b59bb3097> (accessed on 1 May 2026).
6. Koopmanschap, M.A.; Rutten, F.F.; van Ineveld, B.M.; Van Roijen, L. The friction cost method for measuring indirect costs of disease. *J. Health Econ.* **1995**, *14*, 171–189. [CrossRef] [PubMed]
7. Zhang, W.; Bansback, N.; Anis, A.H. Measuring and valuing productivity loss due to poor health: A critical review. *Soc. Sci. Med.* **2011**, *72*, 185–192. [CrossRef] [PubMed]
8. Arredondo, A.; Orozco, E.; Alcalde-Rabanal, J.; Navarro, J.; Azar, A. Challenges on the epidemiological and economic burden of diabetes and hypertension in Mexico. *Rev. Saúde Pública* **2018**, *52*, 23. [CrossRef] [PubMed]
9. Instituto Nacional de Estadística y Geografía (INEGI). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2022, Nueva Serie* [Dataset]; INEGI: Aguascalientes, Mexico, 2023. Available online: <https://www.inegi.org.mx/programas/enigh/nc/2022/> (accessed on 1 May 2025).
10. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **2018**, *21*, C1–C68. [CrossRef]
11. Schultz, T.P. Assessing the productive benefits of nutrition and health: An integrated human capital approach. *J. Econom.* **1997**, *77*, 141–158. [CrossRef]
12. Saha, S. Impact of health on productivity growth in India. *Int. J. Econ. Financ. Manag.* **2013**, *2*, 303–312. [CrossRef] [PubMed]
13. Siddique, H.M.A.; Mohey-ud-din, G.; Kiani, A. Human health and worker productivity: Evidence from middle-income countries. *Int. J. Innov. Creat. Chang.* **2020**, *14*, 523–544.
14. Chapman, K.R.; Mannino, D.M.; Soriano, J.B.; Vermeire, P.A.; Buist, A.S.; Thun, M.J.; Connell, C.; Jemal, A.; Lee, T.A.; Miravittles, M.; et al. Epidemiology and costs of chronic obstructive pulmonary disease. *Eur. Respir. J.* **2006**, *27*, 188–207. [CrossRef] [PubMed]
15. Li, S.; Wang, J.; Zhang, B.; Li, X.; Liu, Y. Diabetes mellitus and cause-specific mortality: A population-based study. *Diabetes Metab. J.* **2019**, *43*, 319–341. [CrossRef] [PubMed]
16. Rozjabek, H.; Fastenau, J.; LaPrade, A.; Sternbach, N. Adult obesity and health-related quality of life, patient activation, work productivity, and weight loss behaviors in the United States. *Diabetes Metab. Syndr. Obes. Targets Ther.* **2020**, *13*, 2049. [CrossRef]
17. Williams, J.; Allen, L.; Wickramasinghe, K.; Mikkelsen, B.; Roberts, N.; Townsend, N. A systematic review of associations between non-communicable diseases and socioeconomic status within low-and lower-middle-income countries. *J. Glob. Health* **2018**, *8*, 020409. [CrossRef]
18. Mayne, T.J.; Howard, K.; Brandt-Rauf, P.W. Measuring and evaluating the effects of disease on workplace productivity. *J. Occup. Environ. Med.* **2004**, *46*, S1–S2. [CrossRef] [PubMed]
19. Mayne, T.; Beresford, T.P.; Overton, J. Absenteeism and presenteeism: Health-related productivity losses associated with depression. *J. Occup. Environ. Med.* **2004**, *46*, 606–611.
20. Rongen, A.; Robroek, S.J.; van Lenthe, F.J.; Burdorf, A. Workplace health promotion: A meta-analysis of effectiveness. *Am. J. Prev. Med.* **2013**, *44*, 406–415. [CrossRef] [PubMed]

21. Proper, K.I.; van Oostrom, S.H. The effectiveness of workplace health promotion interventions on productivity outcomes: A systematic review of systematic reviews. *Syst. Rev.* **2023**, *12*, 45. [[CrossRef](#)] [[PubMed](#)]
22. Song, Z.; Baicker, K. Effect of a workplace wellness program on employee health and economic outcomes: A randomized clinical trial. *JAMA* **2019**, *321*, 1491–1501.
23. Jeet, G.; Thakur, J.S.; Prinja, S.; Singh, M. Community based interventions for non-communicable diseases prevention and control in low- and middle-income countries: A systematic review. *PLoS ONE* **2017**, *12*, e0180640. [[CrossRef](#)] [[PubMed](#)]
24. Wipfli, H.; Zacharias, K.D.; Hundal, N.R.; Reynales Shigematsu, L.M.; Bahl, D.; Arora, M.; Bassi, S.; Kumar, S. Workplace wellness programming in low- and middle-income countries: A qualitative study of corporate key informants in Mexico and India. *Glob. Health* **2018**, *14*, 46. [[CrossRef](#)] [[PubMed](#)]
25. Drummond, M.F.; Sculpher, M.J.; Claxton, K.; Stoddart, G.L.; Torrance, G.W. *Methods for the Economic Evaluation of Health Care Programmes*; Oxford University Press: Oxford, UK, 2015.
26. Domínguez-Miranda, S.A.; Rodríguez-Aguilar, R. The economic dimensions of non-communicable diseases: A panel data study. In *International Conference on Intelligent Computing & Optimization*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 134–144. [[CrossRef](#)]
27. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [[CrossRef](#)]
28. Chen, Y.H.; Sawan, M. Trends and challenges of wearable multimodal technologies for stroke risk prediction. *Sensors* **2021**, *21*, 460. [[CrossRef](#)] [[PubMed](#)]
29. Domínguez-Miranda, S.A.; Rodríguez-Aguilar, R. Related factors with NCD in developing countries: Economic, diet and risk factors dimensions. *EAI Endorsed Trans. Pervasive Health Technol.* **2024**, *10*. [[CrossRef](#)]
30. Domínguez-Miranda, S.A.; Rodríguez-Aguilar, R.; Velázquez-Salazar, M. Modeling the relation between non-communicable diseases and the health habits of the Mexican working population: A hybrid modeling approach. *Mathematics* **2025**, *13*, 959. [[CrossRef](#)]
31. Athey, S.; Imbens, G.W. Recursive partitioning for heterogeneous conditional effects. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 7353–7360. [[CrossRef](#)] [[PubMed](#)]
32. INEGI. *Encuesta Nacional de Ingresos y Gastos de los Hogares 2022: Diseño de Muestra y Factores de Expansión*; Instituto Nacional de Estadística y Geografía: Aguascalientes, Mexico, 2023.
33. INEGI. *Cálculo de Los Índices de Productividad Laboral y Del Costo Unitario de la Mano de Obra*; Instituto Nacional de Estadística y Geografía: Aguascalientes, Mexico, 2015.
34. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
35. Louppe, G. *Understanding Random Forests: From Theory to Practice*; Cornell University Library: Ithaca, NY, USA, 2014.
36. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794. [[CrossRef](#)]
37. INEGI. *Producto Interno Bruto Por Entidad Federativa 2022*; Instituto Nacional de Estadística y Geografía: Aguascalientes, Mexico, 2023.
38. SHCP. *Presupuesto de Egresos de la Federación Para el Ejercicio Fiscal 2022*; Secretaría de Hacienda y Crédito Público, Gobierno de México: Mexico City, Mexico, 2022.
39. SHCP. *Análisis de Egresos: Clasificación Funcional del Gasto (Salud)*; Secretaría de Hacienda y Crédito Público, Gobierno de México: Mexico City, Mexico, 2022.
40. Secretaría del Trabajo y Previsión Social. (2018, 23 de octubre). *Norma Oficial Mexicana NOM-035-STPS-2018, Factores de Riesgo Psicosocial en el Trabajo—Identificación, Análisis y Prevención*. Diario Oficial de la Federación. Available online: <https://asinom.stps.gob.mx/upload/nom/48.pdf> (accessed on 27 July 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.