Project Title: Leveraging Data Science and Machine Learning to Improve Equity in Oversight of H-2A Employers

Which DOL Program?

The H-2A guestworker program addresses the hiring needs of agricultural employers and provides employment opportunities to migrant workers. The Department of Labor (DOL)'s Wage and Hour Division (WHD) enforces essential protections for workers by (1) monitoring employer compliance with the program's terms, such as workers' rights to pay for work performed, safe worksites, and habitable living conditions, (2) investigating noncompliance, and (3) sanctioning employers who are found to be in violation. Our proposal has two aims focused on improving equity in DOL oversight:

- **Aim one: leveraging local data sources to identify potential false negatives in WHD compliance data:** how can DOL combine its federally-held data with local data sources from direct service organizations that might help identify "false negatives" in DOL oversight, such as employers who may be failing to protect underserved workers but who have never been investigated by WHD?
- **Aim two: leveraging external data sources, text as data, and machine learning to predict noncompliance:** are there ways to equitably identify specific establishments that are at a high risk of program noncompliance?

How Do these Aims Promote Equity for Underserved Populations?

These two project goals, and our project's focus on the H-2A guestworker program, align with two sets of agency priorities.

First, Executive Order 13985's focus on the advancement of equity aligns with our focus on oversight of the H-2A program.[1] In FY2020, DOL certified 275,430 H-2A positions and saw an 8% increase in the number of applications (USDA Economic Research Service, 2020). As the Economic Policy Institute summarizes in a recent report on labor violations in the agricultural industry, H-2A workers, who comprise 10% of the agricultural workforce and overwhelmingly migrate from Mexico (approximately 94%),[2] face heightened barriers to reporting employer noncompliance than agricultural workers who are U.S. citizens or permanent residents (Costa et al., 2020). As the report notes, H-2A workers lack "full rights and agency in the labor market," making them "vulnerable to violations of their rights because of their immigration status" (p. 2). DOL, in choosing high-priority areas for upcoming Service Equity Assessments, would do well to consider ways to improve equity within the WHD compliance process and to focus on industries with high concentrations of underserved workers.

Second, Mathematica's June 2020 analysis of Directions for Future Research in WHD Compliance Strategies argues that a key barrier to estimating the prevalence of violations in a given program like H-2A are limitations in the WHISARD data (Dolfin et al., 2020, p. 24). Here, we focus on two of the identified limitations. First are *false negatives,* or establishments that are never investigated but that have substantial compliance issues that harm the rights of underserved populations (Dolfin et al., 2020, p. 24). Important for the Equity EO are *systematic underestimates* of violations due to equity barriers in steps that

[1] For shorthand, we call this the "Equity EO" in the remainder of the discussion.
[2] Department of Homeland Security, 2020.

precede an investigation. For instance, a 2017 GAO report finds that there are many disincentives for workers to report abuse, including fears of retaliation, blacklisting from future work opportunities by employers or farm labor contractors (FLCs), and violence against family members back home if they report issues in their current employment (Government Accountability Office (GAO), 2017, p. 37). If workers fear reporting, or overcome those concerns but struggle with administrative burdens in the reporting process (Herd and Moynihan, 2018), WHD compliance data may also contain "false negatives," or establishments that are never investigated but that have substantial compliance issues that harm the rights of underserved populations.

- **Aim One** discusses how we will leverage a unique, non-DOL data source: establishment-level, timestamped intake data from the Analytics and Research Team at Texas Rio Grande Legal Aid (TRLA), the nation's second largest civil legal aid provider. TRLA also houses Southern Migrant Legal Services, which conducts outreach to and fields intakes from agricultural workers in Louisiana, Arkansas, Mississippi, Alabama, Tennessee, and Kentucky. These data contain alleged violations in the H-2A program for Texas and these six other states. We will use this on-the-ground intake data to investigate (1) the possible presence of false negatives in WHISARD compliance data, and (2) equity questions raised by those false negatives. This analysis complements WHD's own internal efforts to mitigate biases from solely investigating complaints, including investigating nationally-representative samples of employers (GAO, 2017, footnote 116) and strategic enforcement within priority industries (GAO, 2017, p. 51).

Second, Dolfin et al. (2020) also note that WHISARD contains "limited information about the characteristics of the establishments that are investigated" (Dolfin et al., 2020, p. 28) and recommend linkages to external data sources. One key advantage of these linkages is the ability to not only summarize patterns of noncompliance using the limited number of fields internal to WHD compliance data—for instance, using logistic regression, Costa et al. (2020) investigated commodities and counties with higher rates of violations[3]---but also to use supervised machine learning (SML) to try to optimize prediction of compliance issues using a large number of fields from these external linkages.[4]

- **Aim Two** discusses our proposed linkage process for these external data sources and our use of a variety of computational methods---using the unstructured text of the job posting as data; train and test set validation---to predict the presence of compliance issues.

What Data Sources do We Use?

[3] The present project builds upon the EPI's report in three primary ways. First, while the EPI report contains an in-depth analysis of trends over time in labor violations and predictors of those trends, it does not contain large-scale external data linkages beyond the QCEW. Second, we use the unstructured text of the job postings to predict investigations and violations. Third, and focusing on a smaller geographic area---Texas, Kentucky, Alabama, Mississippi, Louisiana, Tennessee, and Arkansas---we investigate whether we can address false negatives in the WHD Compliance Action Data with civil legal aid data on intake activity surrounding the H-2 program, where active outreach makes false negative biases potentially less salient. These seven states, while a subset of all covered by WHD oversight, are both high users of the H-2A program (15% of certified H-2A positions in FY2020 disclosure data) and represent an outsize subset of H-2A WHD violations relative to their share of states. Focusing on the latter, and looking at H-2A violations via WHD investigations, the seven states account for some of the highest-violator states (Mississippi, Kentucky) and together, comprise over 18% of all H-2A violations in the most recent, 04/15/2021, WHD data release.

[4] In particular, SML is useful for cases where the number of predictions exceeds the number of observations. As we discuss later, this becomes relevant for our analyses that use the unstructured text of authorized H-2A job postings to predict the presence of compliance issues.

Table 1 describes the DOL public use data sources and non-DOL data sources we will use. A key advantage of the project are the two data sources highlighted in green, which provide unusually rich insight into (1) establishments that might never appear in that system but that may suffer from compliance issues and (2) textual characteristics of job clearances that predict noncompliance within the WHD oversight system.

Table 1. Data Sources

Data source	**DOL or external?**	**Used for**
WHD Compliance Action Data	DOL public use dataset	One measure of noncompliance, to: (1) Predict whether WHD investigated an employer (2) Among employers investigated, predict whether violation found and CMPs assessed
ETA OFLC H-2A Disclosure Data	DOL public use dataset	Structured dataset of H-2A listings: - Through its total worker fields, provides a "denominator" for the compliance data, or the universe of entities that could have compliance issues - Contains job requirement characteristics like education level, background check and drug screening requirements, pay deductions, and housing types/locations
Scraped daily job postings and unstructured text in Addendum C, "additional material terms and conditions of the job offer"	DOL public use dataset but aggregated by TRLA as part of daily scraping for an LSP-directed job postings site: **https://trla.shinyapps.io/H2Data/**	Complements quarterly disclosure data by: - Daily collection at the time a detailed job order is posted via a State Workforce Agency (SWA) and then aggregated on the Seasonal Jobs website - Including unstructured text data from Addendum C of Clearance Orders that can be preprocessed and used to predict the presence of future compliance issues
ETA OFLC Debarment data	DOL public use dataset	Will be used to construct features such as "ever temporarily debarred"; "number of temporary departments" for predictive model
ACS contextual characteristics of employment and housing locations	External	Additional predictors; we have geocoded all employer locations and job sites in the H-2A jobs data, and can merge tract-level demographic and contextual information
TRLA intake data	External	Employer-level dataset by month and year of alleged noncompliance and other legal issues within the H-2A program

What are the Methods of Analysis and Research Outputs?

- **Step 1: linkages between data sources to construct a dataset where (1) each establishment with any H-2A postings is a row and (2) each establishment is repeated across potentially-active months**
    - Since each data source contains establishment names and locations but different unique identifiers (e.g., H-2A job listings contain Case Numbers; WHD Compliance Action data contains a different Case ID), we will use probabilistic record linkage based on establishment name(s) and address to link the various data sources
- **Step 2: cleaning and constructing three types of "features/predictors" of potential compliance issues**
    - **Structured fields in each data set:** these are more traditional fields like NAICS codes, locations, and the demographic context of the local area
    - **"Bag of words" and other unsupervised feature creation using unstructured text within job postings/addendums:** in addition to standard predictors, machine learning allows us to predict the outcome using high-dimensional data. H-2A orders contain Addendum Cs with free-text fields summarizing "additional material terms and conditions of the job offer." We will first treat the text of these as a "bag of words" (e.g., a job posting with "must be able to work under extreme weather conditions" becomes, after preprocessing to remove common words, the following entry in a "job-term" matrix (Grimmer and Stewart, 2013)).

CASE_NUMBER	extrem	weather	condit
H-300-20241-793909	1	1	1
...	0	1	0

    - **Expert-informed feature creation:** while unsupervised summaries of the clearance orders may be useful for prediction, we will supplement these summaries with human coding of a random sample of Addendums for "red flag" compliance issues. For instance, in the workplace and housing rules section, does the order include very strict workplace and housing rules, stating that the worker can be fired for something minor, like littering, using a cell phone, or taking breaks? If the clearance order contains a productivity standard, is it a facially reasonable one? We will have expert readers flag the Yes/No presence of red-flag signals in a subset of addendums and generalize to the remainder using an imputation model.
- **Step 3: using predictors/features *preceding* an investigation or intake call, predict three compliance outcomes for a given establishment-month:**
    - **Investigation as detected in the WHD compliance data:** this helps us analyze potential compliance issues as reported to DOL WHD;
    - **WHD-detected violation conditional on an investigation:** this helps us analyze confirmed compliance issues;
    - **Investigation as detected in the TRLA intake calls (restricting to establishments in the six states in the TRLA catchment area):** this helps us analyze potential false

negatives in the WHD Compliance Actions by using an alternate source of information on issues.