II. Data

Since the 1970s, the Intelligence Division of DEA has been managing STRIDE, a forensic database containing detailed information on the amount, type, potency, and source of drugs acquired by DEA through their various law enforcement activities. The primary purposes of the database are to control the inventory of drug acquisitions and to assay the characteristics of the drugs collected so as to keep law enforcement informed about the drugs being traded and to assist prosecutors in the prosecution of federal drug offenders. The STRIDE database contains information only on drug acquisitions that are sent to a DEA laboratory for analysis. Thus, the vast majority of the data reflect drug acquisitions obtained by DEA agents or through their informants. To the extent that other agencies rely on DEA laboratories for analysis of their own drug acquisitions (e.g., the Washington DC Metropolitan Police Department), such acquisitions are also included in STRIDE. However, most local and state police agencies do not submit information on their undercover acquisitions to DEA laboratories.

The STRIDE data are not collected for the purpose of conducting statistical analyses of drug transactions. Thus, observations are not obtained through the construction of a probabilistic sampling frame of all drug market transactions in a given geographic area. Instead they represent specific transactions that were targeted by law enforcement agencies. Further, the timing and location of encounters varies considerably from quarter to quarter and from year to year. Therefore, the information about drug transactions provided in the data is not representative of all drug transactions that occur in these areas. That does not mean, however, that these data do not provide information. Instead, it means that the distribution of price and purity values observed in the STRIDE data does not mirror the actual distribution of price and purity observations in the United States, and standard statistical assumptions regarding the asymptotic properties of the distribution of data in STRIDE will not hold. Numerous statistical methods have been developed in recent years to deal with drawing inferences from non-representative data and convenience samples. The current report uses some of these new methods and proposes the adoption of additional methods that could be pursued in future efforts.

Of course, these statistical methods cannot address possible systematic differences between prices paid by law enforcement agents and their informants as recorded in STRIDE, on the one hand, and prices paid by criminal participants in the drug trade, on the other. For example, if law enforcement agents and their informants systematically paid 10 percent more to acquire drugs than did criminal participants, then every observation in STRIDE would be inflated by 10 percent relative to what criminal participants pay, and no analysis based on the STRIDE data alone could detect or adjust for that 10 percent differential. It is not uncommon to assume that law enforcement agents and their informants must pay very nearly the market price; otherwise, the other parties to the transaction would realize who they are. A similar but milder assumption is that law enforcement agents and their confidential informants may pay systematically more (or perhaps less) than do typical market participants, but that any differential is likely to be stable over time and, hence, will affect only estimated price levels, not trends in those levels. Since the present analysis is based only on STRIDE data, we cannot comment on the reasonableness of either of those assumptions.

Acquiring the STRIDE Data

Historically, DEA has been willing to share nonconfidential information in STRIDE with other government agencies and the public. However, no formal codebook of the database exists. Hence, persons interested in using these data have not always known exactly what to ask for or how to ask for it. Equally problematic, non-DEA users may not fully understand what it is that they have requested. For example, early in this project we received a download of STRIDE data that did not contain the FORM field. This field identifies the units in which the transaction amount is measured (e.g., grams, capsules, milliliters). While most observations are measured in grams, about 4 percent of the observations in STRIDE are measured in other units, with specific drugs being affected differently. For example, 16 percent of the methamphetamine observations are measured in units other than grams, while less than 1 percent of the cocaine observations are measured in grams. It is our understanding that some past users of STRIDE data received similar files and believed that all of the observations were measured in grams. Clearly, mistakenly assuming that an observation describes the amount paid for 2 grams when in fact it was the price paid for 2 capsules can affect price estimates, even when the proportion of observations measured in units other than grams is small.

As another example, updated downloads of STRIDE (as opposed to new, complete downloads) can contain nearly but not exactly identical records concerning the same transaction. The problem with simple updates of STRIDE is that cases that are seized but not yet analyzed could show up twice, once with incomplete information for various fields and again with more complete information after the data have been analyzed in the lab. Cleaning the merged (original plus updated) data by eliminating only exact duplicate records would not eliminate such double counting. Double counting of an observation, particularly one that happens to have an unusually high or low price can clearly affect estimates of average prices.

An implication of these kinds of complications is that an important first step in documenting the work done for this project is to describe how the data were requested so that other researchers and analysts can replicate the original starting sample of data. The Office of National Drug Control Policy (ONDCP) requested that a colon-delimited ASCII data file be constructed that satisfied the following three main criteria[1]:

(1) The *date analyzed* (not the date the observation was acquired/seized) must be between January 1, 1981, and "the present" (the letter was dated June 27, 2003).

(2) The following Primary Drug Categories must be included: Cocaine (620), Heroin (610), Cannabis (531), Methamphetamine (111), MDMA & other hallucinogens (560), Heroin signature program (904), Domestic monitoring program (905), CHEMCON (906) and Cocaine signature program (912).

(3) All open and closed cases must be included.

[1] Even the colon delimitation is important. Some past requests have been for semicolon delimited files, but it turns out that (at least) two of the STRIDE fields contain semicolons within the field for some records. Hence, all fields to the right of those fields containing semicolons will report incorrect information.

Requesting information based on the date analyzed instead of the date seized helps reduce the likelihood of getting duplicate observations. Further, information on purity is recorded only after the data have been analyzed. Prior to analysis, the purity field may have missing information or a zero. By limiting the data to only those observations that have been analyzed (i.e., using *date analyzed*), it is possible to interpret zero and missing purity information as real information about the transaction.

It is important to request both open and closed cases to obtain all observations that have been analyzed by the labs. Some fields included in STRIDE are identified by DEA as "sensitive" because they contain information that is pertinent to a case currently being processed by the judicial system (i.e., an open case). If these fields are unwittingly requested, DEA will send information only on closed cases, to protect the integrity of cases currently under investigation. Since the average case is open for about three years, conducting analyses on only closed cases can dramatically limit the number of observations available for evaluating recent trends in prices and/or purity. The typical user is unlikely to know which fields are sensitive and which are not. However, by making a request that both open and closed cases are desired, it is possible for DEA to follow up with the requester to determine whether the sensitive field is truly necessary or not, so that all the data can be sent.

Specific fields requested for this project included the following: city, state, country, date received in lab, date collected, date analyzed, drug name, exhibit number, federal number, lab number, method of acquisition, domestic monitoring program flag, net collected, net weight pure, non-DEA case number, number of packages, office code, collecting office name and code, potency, price per pure gram, primary drug category, primary drug code, received amount, related inventory number, signature, status, STRIDE identifier, submitted amount, total purchase cost, agency submitting, enforcement group, form, suspected drug category, suspected drug code, secondary drug code, dosage, number of doses, number of packages, and package description. Inclusion of the related inventory number is particularly important for cocaine and heroin observations because it enables the user to identify those observations that appear twice in the dataset. Double entries occur for some cases obtained through the Heroin Signature Program (HSP) and the Cocaine Signature Program (CSP) because acquisitions obtained through these two initiatives can be sent to the laboratory twice. When the samples from these initiatives are first sent to the lab, they are entered into the forensic database with special drugcodes indicating that they were obtained through one of the signature programs. They are not given a standard STRIDE ID because the sample is analyzed differently, with the goal of obtaining information about the country or region of origin (i.e., the "signature") of the plant material in the sample. The sample may then later be sent to another lab to examine the purity of the drug it contains, at which point it is given a STRIDE ID and entered into the database a second time. The two entries are linked through the related inventory number.

The file we received from DEA contained 782,031 records, including domestic and foreign data from January 1, 1981, through May 31, 2003.

Constructing the Sample

We began with the raw data file sent to ONDCP and then imposed several restrictions on the data to limit the sample to only the data needed for this analysis. Table 1 summarizes how specific restrictions incrementally affected the sample. For ease of presentation, we use the term *MJ* to refer to marijuana observations, *Meth* to refer to d-methamphetamine observations, and *DMP* to refer to observations from the Domestic Monitoring Program.

(1) Basic Steps to Reduce the Sample to Relevant Observations

We started by identifying, by drug category, all cocaine (drug category 620), heroin (610 and 905), methamphetamine (111), and marijuana (531) observations. The decision to include DMP observations (primary drug category 905) in the sample was made after hearing from DEA personnel that these observations are acquired in a manner consistent with the bulk of the other STRIDE heroin observations. This is not generally true about the observations acquired through the CSP and the HSP, which is why observations from these drug codes were dropped. This restriction reduced the sample to 753,845, as seen in the second row of Table 1. Next, we checked for duplicate records and found none. We then restricted the sample to include only observations from within the United States, which eliminated a little less than 3 percent of the overall sample. An additional 93 observations were dropped because of missing information on the state in which the acquisition occurred. These deletions impacted primarily the sample for heroin.

The next major restriction imposed on the data was to limit the sample to observations where the method of acquisition was either a purchase (P), a seizure (S), or a lab seizure (L). This restriction was based on a recommendation made by DEA, who explained that information obtained from other types of acquisitions (e.g., “flashing money”) may be less reliable than that acquired through purchase attempts and seizures, because such transactions may not have been completed. This restriction reduced the total sample by 5.2 percent, to 693,648, with heroin again having the greatest relative decrease in sample size.

Table 1. Impact of Sample Restrictions on Number of Observations Included in Analysis

	Number of observations remaining in sample							
	Total	**MJ**	**Meth**	**Heroin**	**DMP**	**Cocaine**	**Powder Cocaine**	**Crack Cocaine**
Starting number	782,031							
Restrict to 4 main drug classes	753,845	212,621	88,647	105,021	8,894	338,662		
Only U.S.	731,437	210,897	87,130	97,251	8,514	327,645		
Non-missing state	731,344	210,882	87,130	97,179	8,514	327,639		
Purchases and seizures only	693,648	204,057	81,531	89,123	8,465	310,472		
Raw weight > 0	690,749	203,625	81,233	88,622	8,455	308,814		
Measured in grams	662,168	197,009	68,055	86,542	8,294	302,268		
Purity is non-missing and purity ≤ 100	662,114	197,007	68,052	86,536	8,294	302,225		
Narrowing drug codes	600,139	193,641	60,926	60,558	8,294	276,720	154,155	122,565
Reassigning heroin DMP	599,425	193,641	60,926	68,138			154,155	122,565
Weight ≥ 0.1 gram	554,815	186,637	50,002	63,482			145,353	109,341
Final sample for purity analysis: purchases and seizures	**368,178**	**---**	**50,002**	**63,482**			**145,353**	**109,341**
Price > 0 and non-missing	137,222	4,695	12,313	27,953			45,618	46,643
Remove other gross outliers	136,505	4,597	12,232	27,797			45,423	46,456
Delete crack if year < 1986	136,268	4,597	12,232	27,797			45,423	46,219
Delete obs in city-quarters with < 5 obs.	136,213	4,597	12,181	27,797			45,419	46,219
Stage I: sample for purity models	128,283	---	11,976	27,262			44,913	44,132
Stage II: final sample for price models	131,184	4,359	11,682	26,594			43,953	44,596

* Note that marijuana is not evaluated in the purity analyses with purchases and seizures because data on purity of marijuana are not available in STRIDE.

(2) Preliminary Data Cleaning

The next steps focused on the two primary descriptive variables in the data: amount and potency. First, we deleted observations with missing or zero amounts, where amount represents the raw weight of the purchased or seized package. This had a very small effect on the sample. We then further restricted the sample to observations that were measured in grams, because the remaining forms could not be easily converted into grams. For example, 16.2 percent of the methamphetamine observations were measured in either milliliters (MLS) or capsules (CAP). The precise conversion of these units into grams depends on a number of factors we cannot observe, and hence we decided to delete these observations from the data. This restriction reduced the overall sample available for analysis by 4.1 percent overall. Next, we deleted 54

observations in which either purity was missing or the purity measure was greater than 100 percent. Most of these were cocaine observations.

The next data-cleaning step involved the identification of more homogenous drug codes within specific drug categories. A major criticism raised by the National Research Council regarding past price indices constructed from STRIDE data concerned the aggregation of prices across different forms of a drug that could represent different drug products to consumers.[2] To address this concern, we examined the drug codes represented under specific drug categories and consulted with representatives from DEA to identify those drug forms that were and were not likely to be physically distinguishable to a buyer. Using this information, we aggregated within a drug category drug forms that were physically similar and, hence, likely to be indistinguishable to a buyer.

In the case of cocaine, more than 95 percent of the observations fell under three primary drug code categories: 9041L000 (crack), 9041L005 (powder), and 9041L900 (cocaine, salt undetermined), with the first two categories representing nearly 92 percent of all observations. Crack and powder cocaine are easy to differentiate upon physical inspection, so observations from these two drug categories were separated, as shown in the "Narrowing drugcodes" row of Table 1. DEA informed us that the "salt undetermined" category generally reflects cocaine observations that are too small to analyze chemically. They may or may not have been sold to a buyer as a specific form of cocaine (powder or crack), but the lab technicians were unable to determine the salt attached to the drug, given the time and resources available at the time of analysis. Therefore, the observation was labeled "undetermined." Because this category represents a heterogeneous mix of unidentified cocaine types, we exclude it from further analysis. Excluding the "salt-undetermined" and other cocaine drug forms reduced the total cocaine sample by 8.8 percent.

Most of the heroin observations have one of four drugcodes: heroin hydrochloride (9200.005), heroin base (9200.000), Domestic Monitor Program (9DMP.000), and salt undetermined (9200.900). The breakdown of these heroin observations is as follows:

Type	Frequency	Percent
Heroin base	2,415	2.79
Heroin HCl	58,190	61.33
Heroin, salt undet	25,898	27.30
DMP	8,294	8.74

Again, DEA assisted us in the identification of which drug codes to merge. They informed us that heroin base cannot easily be physically distinguished from heroin hydrochloride (HCl) without chemical analysis, and thus it makes sense to group these two products together. Mexican "black tar" heroin is very easy to physically distinguish from heroin powder, but its physical form does not have a one-to-one correspondence to a distinct chemical form of heroin. The specific chemical form of a black tar sample can be determined by the DEA chemists, but the process can be much more time- and resource-intensive than the process for powder forms of

[2] Manski, Charles F., John V. Pepper, and Carol V. Petrie (eds.), 2001. *Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us*, National Academy Press, Washington, DC.

heroin. Hence, it would not be uncommon for the DEA labs to simply ascertain that a black tar sample was heroin but not to further identify its specific type unless there was a specific need for that more precise determination. Many black tar observations are therefore placed in the *heroin, salt-undetermined* drug form. However, not all salt-undetermined observations are black tar heroin. An observation could get classified as salt-undetermined for other reasons. For example, if the heroin is mixed with many other diluents and adulterants containing salt bases, it could again be prohibitively time-consuming for a chemist to distinguish the salt attached to the heroin from the salts attached to the diluents and adulterants. Alternatively, the sample might be too small. Hence, the salt-undetermined category represents a heterogeneous mix of different forms of heroin. As such, it was decided that this form of heroin would remain separate from the other two forms (heroin base and heroin hydrochloride). Further, we decided not to construct a formal price/purity series for this form, since changes in the series could reflect changes in the makeup of the form included in the salt-undetermined sample rather than real trends in price or purity for a particular form of heroin. We verified, however, that a series for the heroin salt-undetermined category follows a different general trend from that observed for the main heroin series (see Section VIII, Supplemental Analyses). For example, in recent years, the purity of salt-undetermined heroin is generally lower than that of heroin base or heroin hydrochloride.

The DMP category consists of purchases made through the Domestic Monitoring Program, a program in which law enforcement goes into specific cities and makes small buys (usually $100) of whatever type of heroin is available on the streets. Thus, these DMP observations are a heterogeneous mix of different forms of heroin. However, the specific forms can be identified through the secondary drug code, and thus it is possible to identify which DMP observations are heroin hydrochloride and which are heroin base. Examination of the secondary drug code for the sample of DMP observations in this dataset revealed that 7,580 observations could be included in the main heroin sample because they were either heroin hydrochloride or base. All other heroin observations were dropped.

Three main types of methamphetamine are marketed: d-methamphetamine, dl-methamphetamine, and l-methamphetamine. These types differ in the form of the isomer, something that is not immediately apparent to the buyer at the time of the transaction. However, according to DEA personnel, the three types of methamphetamine differ significantly in their quality, so sellers usually make the type known to the buyer as a way of indicating the quality of the drug. In the STRIDE data, the great majority of observations measured in grams were of d-methamphetamine. The other two forms of methamphetamine together made up only 10.5 percent of the total methamphetamine sample. Hence, it was possible to develop a price series only for the d-methamphetamine type.[3]

More than 97 percent of the marijuana observations fell into one of two drugcode categories: 7600.000 (no plant material detected) and 7360.4 (all plant material).[4] The next two largest forms, intact plants (7360.5) and cannabis seeds (7360.0), would be very easy to physically differentiate from general plant material and were too small to generate their own price series, so

[3] STRIDE contains numerous drug codes that represent different physical forms of d-methamphetamine. According to information from DEA personnel, the physical form is not nearly as important to a buyer as the type. Thus, for the construction of this methamphetamine price series, all forms of d-methamphetamine were included.
[4] These two forms were grouped together on the basis of a recommendation by DEA.

they were dropped. Observations from the drug code labeled Tetrahydrocannabinol-Organic (7371.000) were retained, however, because this category indicated a plant-like material, which would be difficult to physically distinguish from the other forms. Limiting the marijuana sample to these three drug codes decreased the sample by only 1.7 percent.

(3) Gross Outlier Deletions

Next, we deleted observations with a weight of less than 0.1 grams. This data-cleaning step has been used in previous reports and also by other researchers using the data.[5] The primary justification for the deletion is that purity data are unreliable for observations weighing less than 0.1 grams. As evidence, there is a disproportionate number of low and zero-purity observations at these very small quantities, presumably because it is difficult for lab technicians to chemically determine the potency and specific drug forms involved.[6] Approximately 7.0 percent of the overall sample was lost after imposing this restriction, the largest losses occurring for the d-methamphetamine sample (18 percent) and the crack cocaine sample (11 percent).

Excluding the marijuana observations because they contain no information on purity, the resulting sample of 368,178 observations was the primary sample used to evaluate purity with both the seizure and purchase data. Additional data-cleaning steps were taken to arrive at the sample for price and purity models that employed primarily purchase data.

The first step taken to generate a sample for estimating price, after bringing marijuana observations back into the sample, involved deleting those observations that were missing information on cost. The vast majority of observations with missing price information are seizures, although a few purchase observations are also missing price information.[7] Deleting observations in which price was missing significantly reduced the sample, to 137,222, a 75.3 percent decrease overall. The largest percentage decrease was for marijuana, the sample for which was decreased by 97.5 percent, to 4,695 observations.[8]

Next, observations that were outside the distribution of realistic prices for 1 gram not adjusted for purity were deleted. Criteria for deleting specific data points generally followed those employed in previous reports (see Table 2). For example, observations in which the nominal price (i.e., price not adjusted for inflation) was too low or the inflation-adjusted (or real) price per gram was too high were dropped. However, in contrast to previous reports, observations in which the inflation-adjusted (real) price per gram was too low were also dropped. This additional restriction was placed on the data following close examination of the distribution of

---

[5] Caulkins, J.P. (1994), *Developing Price Series for Cocaine*, MR-317-DPRC, RAND, Santa Monica, CA; Saffer H. and F.J. Chaloupka (1995). "The Demand for Illicit Drugs" National Bureau of Economic Research Working Paper #5238, August 1995; Grossman M., F.J.Chaloupka, and C.C. Brown (1998), "The Demand for Cocaine by Young Adults: A Rational Addiction Approach," *Journal of Health Economics*, Vol. 17, No. 4, pp. 427–474.

[6] In previous reports, observations for marijuana in which the amount, not adjusted for purity, was less than 0.2 grams were deleted. The methods for marijuana employed here are therefore less restrictive than those used in previous reports.

[7] Crack cocaine purchases had the most cases of missing price information, with 1944 purchase observations missing price. The number of purchase observations missing information on price for powder, heroin, methamphetamine, and marijuana were 1645, 598, 365 and 775, respectively.

[8] There were a few seizure observations that included a non-zero price (0.3% of our sample). Given that price was reported for these observations, they were retained under the presumption that a coding error had occurred labeling these observations as seizures when they were, in fact, purchases.

real prices per gram (not adjusted for purity) for each drug. In total, these conditions reduced the total sample by only 0.5 percent overall. Drug-specific samples were reduced by less than 1 percent, with the exception of marijuana, which had a 2.1 percent reduction.

Table 2. Criteria for Deleting Data Points, by Drug

Drug	Nominal Price	Real Price per Gram	Real Price per Gram
Cocaine	< $ 3.00	< $ 2.00	> $ 3,000
Heroin	< $ 3.00	< $ 7.50	> $10,000
d-Methamphetamine	< $ 3.00	< $ 2.00	> $ 3,000
Marijuana	< $ 0.10	< $ 0.05	> $ 100

(4) Additional Data Cleaning Related to Proper Model Estimation

Because data for crack cocaine were sparse prior to 1986, all crack cocaine observations between 1981 and 1985 were deleted. There were only 137 crack observations for the entire 1981–1985 time period, so runs that were done including these observations generated extremely volatile trends. The volatility was attributed to the enormous sampling error and did not represent true volatility in crack cocaine prices.

Similar artificial volatility was possible in all of the drug models because of the geographic and temporal variability of law enforcement activities. Thus, to limit the amount of artificial volatility in the time series, we required, for each drug model, that a cell, defined as a particular quarter and year, had to have at least five observations to be included in the model. For the marijuana models, at least five observations had to exist for a given year, because the models are estimated on the basis of years. The justification for this restriction was that it would reduce the sensitivity of the model to unusual or outrageously large/small prices observed in specific locations that might otherwise heavily influence the prediction for that quarter-year. Requiring at least five observations in a given quarter increases the probability that deleting observations with extreme residuals (described below) will delete the unrealistically priced observations. This restriction deleted only 51 methamphetamine and four powder cocaine observations.

The resulting sample of 136,213 observations was the starting sample for the estimation of the price/purity models.

As will be described in Chapter IV, the first stage of the price/purity model—the purity equation—excluded observations with a purity of zero. The reason for this is that a zero purity in STRIDE may indicate a true zero purity or it may indicate that the purity was missing (not yet determined). Thus, to avoid incorrectly assigning zero to the purity of many observations, we just deleted these observations. This reduced the samples for d-methamphetamine, heroin, and cocaine powder by less than 2 percent, but it reduced the crack cocaine sample by 4.5 percent. Once the purity model is run, however, it is possible to generate a predicted purity for all of these zero-purity observations, allowing them to be brought back into the analysis for estimating the final price model. Thus, the omission of zero-purity observations influenced only the sample of observations predicting the purity model.

The final step to reach the end sample for the price model was the exclusion of extreme residuals. The exact process for deleting extreme residuals (described in Chapter IV) basically involved an iterative loop where extreme residual observations were dropped from the model until no more could be identified. Comparing the last row in Table 1 to the third-from-last row, one can see the impact of deleting the extreme residuals from the samples. Overall, the total sample decreased by 3.7 percent, with the sample for marijuana being the most impacted. The final total data sample contained 131,184 observations.

(5) Other Modifications to the Data

In addition to the above general modifications, some additional issues needed to be addressed. First, while this was not an issue for the last data extract from DEA, a prior extract from STRIDE had duplicate records, as indicated by the STRIDE ID variable. Duplicate records can occur if users request updates of the STRIDE database instead of complete downloads starting at the same base year, because cases in STRIDE that have not yet been analyzed at the lab can be analyzed during the updating period. Both observations would have the same STRIDE ID but would contain different information for potency and other variables of interest. In addition, duplicate records may result from the inclusion of observations from specific operations (e.g., DMP or CSP). These can be identified through a non-missing inventory number, which will specify an original STRIDE ID if one was previously assigned to that observation. Finally, some records in STRIDE appear to look like duplicates, although different STRIDE IDs have been assigned or specific variables (e.g., lab number) differ slightly. Although these "nearly duplicate" records are maintained in the current analysis, future work should evaluate whether they are indeed unique.

A relatively minor data-cleaning issue was that of the correction of erroneous date codes. A very small number of observations (21) had seizure or purchase dates that were just not possible (e.g., February 29 in a non-leap year or 31 in months with only 30 days). It was presumed that these erroneous dates were coding errors specific to the day of the month, so these observations were back-coded to the closest earlier logical date, following an algorithm employed in earlier reports.

The data-cleaning steps outlined above generally followed those undertaken in previous reports, with three primary differences. First, previous reports aggregated all the drug codes within specific drug categories, so, for example, crack and powder cocaine (as well as the other, less frequent drug codes under cocaine) were estimated together as a single series. Likewise, all drug codes for heroin, d-methamphetamine, and marijuana were included in the samples employed in previous reports. Hence, the series presented in the current main report do not represent the same drugs as the series in previous reports. Second, previous reports did not delete observations in which the real price per gram (not adjusted for purity) was too low. Instead, those reports imposed restrictions that required potency to be above specific thresholds for each drug. The present report does not include this restriction because low-potency observations are viewed as valid "rip offs" in the data, and the model is modified to accommodate the information revealed from these transactions. Third, previous reports did not exclude quarters with too few observations. This additional restriction was imposed in order to reduce artificial volatility in the price and purity trends.

Defining Key Variables

The Dependent Variables

The primary dependent variable in the price model is the inflation-adjusted price (or "real price") of a particular drug. We adjust for inflation using the quarterly Consumer Price Index for All Urban Consumers (1982–1984 base year).[9] All prices are reported in constant 2002 dollars because 2002 is the last year for which the STRIDE data are fully reported. Thus, the prices for observations up to the second quarter of 2002 are adjusted upward for inflation, while those starting in the third quarter of 2002 are adjusted downward.

For the powder cocaine, crack, heroin, and d-methamphetamine price models, a two-stage estimation technique is employed in which the purity of the drug is estimated in a first stage regression and the predicted value of those purity observations, or the expected purity, is put into the right-hand side of the price model. Thus, the purity of each drug is also a dependent variable. Because the unit for potency in the raw data from DEA is *percent* (so something that is 95 percent pure has a value of 95), this variable is converted into a fraction by multiplying each value by 0.01 before entering it into the model. Hence, all of the potencies reported in the paper are presented as fractions, where 1.00 = 100 percent pure.

The Independent Variables

One of the key dependent variables in both the price and purity equations is the amount of the drug involved in the transaction. DEA reports the weight involved in a transaction under its AMOUNT variable. Note that the measure AMOUNT does not adjust for purity; it simply reflects the weight of the transaction in grams.

All of the price and purity models also account for the quarter and year in which a transaction took place. Information on the date of the transaction is obtained from the variable DATE SEIZED. From this we construct a series of quarter/year interaction terms by constructing 90 dichotomous indicators (T1 through T90) representing specific quarter/years from the first quarter of 1981 through the second quarter of 2003. A transaction that took place on March 24, 1990, for example, would have a value of one for the time indicator T37 and zero for all the other time indicators (T1 through T90).

Finally, the models include information on the location where the transaction took place. Previous models estimating price and purity indices using these data included dichotomous indicators for 29 metropolitan areas (called *cities*) in the data and an "other" category that encompassed the rest of the country outside those 29 cities. The cities that have unique identifiers are Atlanta, Baltimore, Boston, Buffalo, Chicago, Cleveland, Dallas, Denver, Detroit, Houston, Kansas City, Los Angeles, Miami, Milwaukee, Minneapolis-St. Paul, New Orleans, New York, Newark, Philadelphia, Phoenix, Pittsburgh, Portland, San Antonio, San Diego, San Francisco, Seattle, Saint Louis, Tampa, and Washington DC.[10]

[9] Historical information on the annual Consumer Price Index for All Urban Consumers (seasonally adjusted) can be obtained from the Bureau of Labor Statistics at http://www.bls.gov/cpi/home.htm.

[10] Other cities in the United States are also represented in the data, and unique identifiers could be included for those with sufficiently large numbers of observations over time.

The current specification of the price model includes dichotomous indicators for the same 29 locations as in the previous report, but the remaining "other" category is subdivided into nine separate Census divisions: Pacific, Mountain, North West Central, East North Central, West South Central, East South Central, South Atlantic, Mid Atlantic, and New England. While the earlier approach would group, say, rural Florida and Montana into the same "other" category, this approach separates these locations by assigning them to their specific region, thus reducing the amount of unexplained variation remaining in prices. The current model identifies 38 geographically distinct areas, whereas previous methods identified only 30.

The nine Census divisions used in this analysis are defined as follows:

- Pacific: Alaska, Hawaii, Washington, Oregon, California
- Mountain: Arizona, Idaho, Montana, Colorado, New Mexico, Utah, Nevada, Wyoming
- West North Central: North Dakota, South Dakota, Nebraska, Kansas, Minnesota, Iowa, Missouri
- East North Central: Wisconsin, Illinois, Michigan, Indiana, Ohio
- West South Central: Texas, Oklahoma, Louisiana, Arkansas
- East South Central: Mississippi, Alabama, Kentucky, Tennessee
- South Atlantic: Florida, Georgia, South Carolina, North Carolina, Virginia, District of Columbia, Maryland, Delaware, West Virginia,
- Mid Atlantic: New Jersey, New York, Pennsylvania
- New England: Connecticut, Rhode Island, Massachusetts, Vermont, New Hampshire, Maine

It is also important to clarify the way the cities are defined. Previous reports tried to aggregate cities into metropolitan statistical areas (MSAs). Definitions for these MSAs, however, have changed over time as new cities and areas have become incorporated. We decided that the current project would employ the most recent definitions from the Office of Management and Budget (OMB) in 2003 to deal appropriately with new locations not previously included in earlier analyses. Using lists of all towns and cities included in specific counties from the Office of Social and Economic Data Analysis (<http://oseda.missouri.edu/plue/>) and MSA definitions based on counties and cities, we constructed SAS code that sorted all of the city names provided in STRIDE into one of the 29 MSAs or the nine Census regions.[11]

[11] A few cities within an MSA did not properly get categorized into that MSA based on the data from the Office of Social and Economic Data Analysis. In these cases, code from the previous report was used to assign these cities to an MSA.