

Issues in using state hospital discharge data in injury control research and surveillance

Bruce A. Lawrence^{a,*}, Ted R. Miller^a, Harold B. Weiss^b, Rebecca S. Spicer^a

^a Pacific Institute for Research and Evaluation, Calverton, MD, United States

^b Center for Injury Research and Control, University of Pittsburgh, Pittsburgh, PA, United States

Received 22 August 2005; received in revised form 3 August 2006; accepted 10 August 2006

Abstract

This study evaluates the quality of injury-related coding in state hospital data and their usefulness to injury researchers. Using 1997 hospital discharge records from 19 states, hospitalized non-fatal injury-related cases were identified by first selecting all cases that met broad criteria for injury, and then dropping cases that appeared incorrectly coded as injuries and cases related to medical care. Based on our criteria, 1,129,980 non-fatal hospitalized cases were identified as probable acute injuries. Three-quarters were coded with a traditional injury diagnosis in the primary diagnosis field, and 90% had a traditional injury diagnosis somewhere in the first six diagnosis fields. Of cases with an injury diagnosis code in the first three diagnosis fields, 88.1% were E coded. E coding completeness varied by state, with some states reporting high rates of E coding by using non-specific E codes. Other challenges included E-coded cases where no injury diagnosis was reported and apparent miscoding of the E code. We conclude that it is possible to combine multiple states' data if researchers are aware of the challenges they may encounter. In order to capture all injury-related cases, it is important to scan secondary diagnosis fields.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Injury; Hospital discharge data; Diagnosis; E codes

1. Introduction

State hospital discharge data (HDD) include records of all hospital discharges in a given year from all acute-care hospitals in the state and are a valuable resource for better understanding factors that increase the risk of or the harm done by injury. Organizations in more than 70% of the U.S. states collect hospital discharge data, and most of them make portions of their data available to health researchers. Just a decade ago state HDD were not very useful to injury researchers because of incomplete or missing codes for the external cause of the injury and poisoning (E codes). E codes are an important tool in designing, implementing, and evaluating injury control programs. Prior to the routine use of E codes, researchers using HDD could only identify the nature and body region of the injury. For example, Mayer et al. (1981) reported that the primary "causes" of severe pediatric trauma were neurological, chest, and extremity

injuries. There was no information on the circumstances that caused these injuries, nor clues on how to prevent or minimize the harm done by them. In 1989 the U.S. Centers for Disease Control and Prevention stated that E coding would be "one of the most effective and feasible means available to collect data needed to prevent and control injuries" (Sniezek et al., 1989). Today, state E coding practices range from legally required entry of an E code in its own distinct field (or fields) to voluntary inclusion of E codes among the diagnosis codes. Organizations that collect hospital data in over half of the U.S. states mandate E codes on hospital discharge records or obtain voluntary E coding compliance exceeding 85%. However, CDC and Medicare uniform billing recommendations for a separate field for E codes have not been implemented by most states.

The usefulness of hospital data to researchers is also affected by the quality of coding, which varies widely from state to state. Data users must be familiar with the peculiarities and pitfalls of a given state's data in order to use it most effectively. E coding quality on several different state hospital discharge datasets has been evaluated (Guyer et al., 1990; Langlois et al., 1995; LeMier et al., 2001; Marganitt et al., 1990; Muelleman et al., 1993; Smith et al., 1990). However, with many states only recently

* Corresponding author at: 11720 Beltsville Drive, Suite 900, Beltsville, MD 20705-3166, United States. Tel.: +1 301 755 2731; fax: +1 301 755 2799.

E-mail address: lawrence@pire.org (B.A. Lawrence).

mandating E coding, quality has not yet been adequately studied. In addition, injury identification is not always straightforward, and little research has investigated the best methods for detecting an injury case using both E codes and diagnosis codes in HDD. Few researchers have combined or compared data across states, possibly because states differ in their data collection policies and procedures, codebooks, and data quality.

The State and Territorial Injury Prevention Directors' Association (STIPDA) recommends for surveillance purposes choosing only the first-listed diagnosis code to identify injury case (Injury Surveillance Workgroup, 2003). Frequently, however, cases E coded as injuries list an injury diagnosis only in a secondary field. We assessed how extensively the STIPDA case definition undercounts injuries.

We constructed an E-coded 1997 injury dataset drawn from 19 states, representing more than half the U.S. population. This article shares our experience throughout this process, the obstacles encountered, and techniques used to resolve problems. It examines the quality of injury-related coding in state hospital data and usefulness of the data to injury researchers. Finally, we make recommendations to both administrators and users of these data.

2. Methods

To compile and aggregate hospitalized injuries, we began with 1997 hospital discharge records from 19 states (Arizona, California, Florida, Maine, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, South Carolina, Utah, Vermont, Virginia, Washington, and Wisconsin). The datasets were obtained from a variety of sources, depending on the state, including state hospital associations, state health departments, and other state government agencies. The methods and findings reported here are an outgrowth of a broader project that combined mortality and hospital discharge data in order to analyze injury patterns and case-fatality rates above an admitted or fatal injury severity threshold. We analyzed data from 1997 because we obtained a large number state hospital discharge files from this year and because 1997 was one of the last years that the mortality data and hospital discharge data systems used the same version of the International Classification of Diseases (the ninth edition, ICD-9). After 1998, mortality data were coded with the 10th edition ICD while hospital data remained coded with ICD-9, Clinical Modification U.S. Department of Health and Human Services (1991).

Together, the 19 state files included 17.8 million records. The ICD-9 diagnoses and E codes used to identify a potential injury record included the following traditional injury codes: (a) 800–994 injury and poisoning (except late effect of complications of surgical and medical care [909.3], late effect of adverse effect of drug, medicinal or biological substance [909.5]) and (b) E800–E999 external causes of injury and poisoning (except location [E849], second-hand tobacco smoke [E869.4]; misadventures to patients during surgical and medical care [E870–E876]; surgical and medical procedures as the cause of abnormal reaction of patient or later complication, without mention of misadventure at the time of procedure [E878–E879]; drugs, medicinal,

and biological substances causing adverse effects in therapeutic use [E930–E949]).

We also included records with the following codes: (a) 363.31 solar retinopathy; (b) 370.24 photokeratitis; (c) 371.82 corneal disorder due to contact lens; (d) 388.11 acoustic trauma (explosive) to ear; (e) 760.5 maternal injury affecting fetus or newborn; (f) 995.5 child maltreatment; and (g) 995.80–995.85 adult maltreatment.

Conceptually, we defined an injury as any ill effect that results from trauma or poisoning unrelated to medical care. Operationally, we defined injury through a complex multi-stage process, first selecting all cases that met any of our broadest criteria for injury, and then whittling it down by dropping cases coded incorrectly as injuries and cases related to medical care. For the final dataset, a primary injury diagnosis and primary injury E code were selected for each case.

Most of the variables were converted to standard formats so that they could be combined into a multi-state injury dataset. In processing the state hospital data, we reconfigured ICD-9-CM codes into two different sets of fields—diagnoses (up to 15) and E codes (up to 6, including a field reserved for location codes).

To identify injury-related cases, we did not rely on just the principal diagnosis (the first diagnosis listed in the record). According to ICD-9-CM coding rules, an E code cannot be used in the principal diagnosis field. Relying on the principal diagnosis field would miss some injury cases, while capturing some non-injury cases. For example, a patient who was admitted for an injury might contract an infection more serious than the injury itself, displacing the injury diagnosis to a secondary field. On the other hand, a patient who was admitted for an illness might suffer a medical misadventure or an adverse drug reaction that results in an injury or poisoning diagnosis in the principal diagnosis field. Therefore, in the first stage of injury selection we kept any case with an injury code in the first three diagnosis fields or an injury E code in any field.

To assign a *primary injury diagnosis* for each injury case, we searched all diagnosis fields and chose the first-listed injury diagnosis. For the cases with an injury E code but no injury diagnosis, we attempted to assign a primary “injury-related” diagnosis from diagnoses 001–799—diagnoses typically classified as illnesses, but which sometimes describe conditions that could be considered injury-related.

Some of the cases selected in the first cut did not represent injury admissions. These cases were dropped from the final dataset of probable injuries. Three categories of cases were scrutinized: (1) those lacking an injury diagnosis; (2) those lacking an injury E code; and (3) those with often-abused E codes (falls, overexertion, unspecified). We examined and classified each category, looking for patterns in the diagnoses and E codes that corresponded to non-injuries, and then deleted the non-injury admissions from the final dataset.

We deleted duplicate records when we found them. From the injury subset thus selected, we made a further cut, dropping all cases that we identified as non-acute, including admissions for late effects, chronic conditions, and rehabilitation. The final combined dataset included probable acute injuries from all 19 states.

Table 1
Numbers of hospital discharge records and identified probable acute injuries, based on hospital discharge data from 19 states, 1997

State	Hospital discharge records	Probable acute injuries	Probable injury percentage	State's share of		Percent of probable acute injuries E coded
				Discharge records	Probable acute injuries	
AZ	516,064	36,309	7.04%	2.90%	3.21%	88.3%
CA	3,685,706	241,722	6.56%	20.71%	21.39%	91.3%
FL	2,031,916	125,199	6.16%	11.42%	11.08%	65.8%
MA	764,635	48,923	6.40%	4.30%	4.33%	95.4%
MD	610,343	41,564	6.81%	3.43%	3.68%	96.2%
ME ^a	156,222	11,036	7.06%	0.88%	0.98%	75.8%
MI ^a	1,242,585	72,639	5.85%	6.98%	6.43%	69.2%
NE	167,394	10,607	6.34%	0.94%	0.94%	93.1%
NH	116,897	8,093	6.92%	0.66%	0.72%	95.0%
NJ	1,426,877	85,424	5.99%	8.02%	7.56%	90.5%
NY	2,446,522	143,947	5.88%	13.75%	12.74%	99.1%
PA	1,821,746	127,433	7.00%	10.24%	11.28%	92.4%
RI	124,855	7,839	6.28%	0.70%	0.69%	96.4%
SC	463,513	24,965	5.39%	2.60%	2.21%	87.4%
UT	225,465	14,175	6.29%	1.27%	1.25%	85.9%
VA	791,968	47,256	5.97%	4.45%	4.18%	77.3%
VT	58,096	4,171	7.18%	0.33%	0.37%	93.4%
WA	536,718	34,218	6.38%	3.02%	3.03%	97.4%
WI	610,962	44,460	7.28%	3.43%	3.93%	98.8%
Total	17,798,484	1,129,980	6.35%	100.00%	100.00%	88.1%

^a E coding not mandatory.

3. Results

Of the 17.8 million records in the combined 19-state file, 1,218,210 (6.8%) met our criteria for selection as a likely acute or non-acute injury, as described above. Of these, 1,129,980 (6.3%) were classified as probable acute injuries (Table 1).

The quantity and quality of hospital discharge data varied markedly from state to state. States provided from 7 to 15 diagnosis fields. A typical state reported 9 or 10 diagnoses and 6 to 10 procedures. Except for California and Washington, all states mix E codes with diagnoses and supplementary (V) codes in the diagnosis fields. (California provided five dedicated fields for E codes. Washington reported a single E code.) There was evidence that, in the absence of a dedicated E-code field, E codes were sometimes crowded out when the limited number of available fields was filled with diagnosis codes. Among injury cases in the Rhode Island data, with only seven diagnosis fields and no dedicated E-code field, nearly half of cases without an E code had all seven diagnosis fields filled. By contrast, in states with 9 or 10 diagnosis fields but no dedicated E-code field (MA, NE, PA, UT, VA, WI, FL, ME, VT), just 17% of cases without an E code had all diagnosis fields filled.

Except for Maine and Michigan, all states whose data we examined in this article had mandatory E coding of injuries. Maine attained a high level of E coding on a voluntary basis. Overall, 88.1% of probable injury cases were E coded (Table 1). Actual E-coding rates varied widely among states, ranging from 99% of injuries in New York and Wisconsin to less than 70% in Florida and Michigan. Twelve of the 19 states exceeded 90%.

Some states attained their high E-coding rates by heavy use of non-specific E codes, such as E928.9 (unspecified accident). Both Maryland and New York, with 96% and 99% E-coding

rates, respectively, used non-specific E codes on 7% of their injury discharges (see Table 2). Seven states had rates of specific E coding greater than 90%, led by Wisconsin, with 93%. Conversely, the three states with the lowest overall E-coding rates, Florida, Maine, and Michigan, showed the least reliance on non-specific E codes. Diagnosis coding, compared to E coding, was much more complete and much more specific (see Table 2). Only 1.1% relied on non-specific diagnoses such as 959 (injury, other, and unspecified).

Most (78.6%) of the 1,129,980 discharge records we identified as probable acute injuries had an injury diagnosis (as defined above) in the principal diagnosis slot, the first diagnosis listed on the hospital discharge record. Among the other 21.4% of probable acute injuries, every major diagnosis range in ICD-9-CM was represented (see Table 3). Fractures of the hip and lower limb accounted for 25.6% of principal diagnoses, other fractures for 17.1%, and other diagnoses in the traditional injury range (800–994) for 35.2%. The most common non-injury principal diagnoses associated with injury-related admissions were mental disorders (4.1%) and circulatory conditions (3.4%).

The first-listed diagnosis, according to NCHS coding rules, should be the diagnosis that primarily caused the patient to be admitted, but states often reserve this position for the diagnosis that contributed most to the length of stay. By diagnostic class, Table 3 shows the diagnoses that most frequently were listed first in probable acute injury cases. The non-injury diagnoses seem to split into post-injury complications and medical conditions that may have precipitated or been discovered in tests performed when treating an injury. Among the probable complications are septicemia, volume depletion, pneumonia, early or threatened labor, cellulitis and abscess, angioneurotic edema, and complications of medical care. Conditions that may have precipitated

Table 2
ICD-9-CM coding of hospital-admitted probable acute injuries, based on hospital discharge data from 19 states, 1997

State	Primary injury diagnosis			Primary injury E code		
	Specific	Non-spec	Missing	Specific	Non-spec	Missing
AZ	98.5%	1.1%	0.5%	85.0%	3.3%	11.7%
CA	98.1%	1.3%	0.6%	86.6%	4.7%	8.7%
FL	98.4%	1.0%	0.6%	64.0%	1.8%	34.2%
MA	98.3%	0.8%	0.9%	90.3%	5.1%	4.6%
MD	97.6%	1.5%	0.9%	89.2%	7.0%	3.8%
ME ^a	98.0%	1.4%	0.6%	74.1%	1.6%	24.2%
MI ^{a,b}	98.9%	1.1%	0.0%	66.7%	2.5%	30.8%
NE	98.4%	0.9%	0.7%	89.1%	4.0%	6.9%
NH	98.3%	1.1%	0.5%	91.4%	3.6%	5.0%
NJ	98.2%	1.0%	0.8%	78.6%	11.9%	9.5%
NY	98.3%	1.3%	0.4%	92.3%	6.8%	0.9%
PA	98.1%	1.1%	0.8%	86.9%	5.4%	7.6%
RI	98.3%	0.6%	1.1%	92.7%	3.7%	3.6%
SC	98.9%	0.9%	0.1%	82.0%	5.4%	12.6%
UT	97.9%	0.8%	1.3%	83.1%	2.8%	14.1%
VA	98.3%	0.9%	0.8%	74.3%	3.0%	22.7%
VT	98.5%	1.0%	0.5%	90.4%	3.0%	6.6%
WA	98.7%	0.5%	0.8%	92.0%	5.4%	2.6%
WI	98.3%	1.0%	0.7%	93.3%	5.5%	1.2%
Total	98.3%	1.1%	0.6%	82.9%	5.1%	11.9%

Non-spec: use of non-specific E codes, such as E928.9 (unspecified accident).

^a E coding not mandatory. The voluntary E coding rate in Maine rose to 90% in 1999.

^b Data provided to for this study included only pre-selected cases with injury diagnoses.

or been discovered because of an injury include diabetes, neoplasms and related pathological fractures, major depression (frequently associated with suicide cases), Alzheimer's, and probably atrial fibrillation, heart failure, and intestinal obstructions (although these events occasionally might be precipitated

by an injury rather than causing an injurious fall). An important question in these cases, which cannot be answered from a discharge record, is whether the injury was serious enough that it would have warranted admission absent the comorbid condition. Finally, syncope and collapse and epilepsy in traumatic brain injury cases could either be causes of injury or complications.

With the principal diagnosis field so often occupied by non-injury diagnosis codes, the primary injury diagnosis often had to be taken from a secondary diagnosis field. An injury diagnosis was listed first in 78% of cases we identified as probably acute injuries. In another 12%, the injury diagnosis was listed in the second or third field. The large majority (95%) of the injury cases had an injury diagnosis in the first six diagnosis fields and 97% had one in the first nine fields.

Overall, in 97.5% of the probable acute injury records a traditional injury diagnosis (800–994) was identified in at least one diagnosis field. The remaining 2.5% of cases contained an injury E code but not a traditional injury diagnosis. The illness diagnoses assigned most frequently to these cases in lieu of traditional injury diagnoses were musculoskeletal conditions (0.7%), adverse effects (0.3%), and skin conditions (0.3%). In 0.6% of cases overall, we were unable to assign any primary injury diagnosis.

A common data problem that made identification of injuries a challenge was the use of an E code in cases where no injury had occurred. The two most frequently misused categories of E codes were falls (E880s) and overexertion (E927). For example, if someone falls after a stroke, some coders will erroneously code the fall, even if no injury occurs. And some coders will attribute a heart attack to overexertion, even though this code is intended to be used with musculoskeletal injuries like strains and sprains. We excluded these non-injury admissions.

Table 3
Principal diagnosis reported on the hospital discharge record of probable acute injuries from 19 states, 1997

ICD-9-CM	Diagnosis description	Records	Percent	Most common condition in category
001–139	Infectious/parasitic disease	5,310	0.5%	Septicemia
140–239	Neoplasms	5,661	0.5%	Lung cancer
240–289	Endocrine/nutrition/metabolic/immunity/blood	14,738	1.3%	Volume depletion/diabetes
290–319	Mental disorders	46,390	4.1%	Major depression
320–389	Nervous system and sense organs	7,245	0.6%	Epilepsy/Alzheimer's
390–459	Circulatory system	38,051	3.4%	Heart failure/atrial fibrillation
460–519	Respiratory system	24,240	2.1%	Pneumonia
520–579	Digestive system	12,066	1.1%	Intestinal obstruction
580–629	Genitourinary system	6,208	0.5%	Urinary tract infection
630–676	Complications of pregnancy/childbirth/puerperium	4,832	0.4%	Early or threatened labor
680–709	Skin and subcutaneous tissue	18,903	1.7%	Cellulitis and abscess
710–739	Musculoskeletal system and connective tissue	20,272	1.8%	Pathological fracture
740–779	Congenital anomalies and perinatal conditions	484	0.0%	Spine/respiratory
780–799	Symptoms, signs, ill-defined conditions	27,691	2.5%	Syncope and collapse
800–994	Injury and poisoning	880,173	77.9%	Hip fracture/other fracture
995	Adverse effects	2,522	0.2%	Angioneurotic edema
996–999	Complications of surgical/medical care	4,853	0.4%	Mech complication of implant
V01–V82	Factors influencing health status	10,328	0.9%	Observation
	Missing	13	0.0%	
		1,129,980		

Medical misadventures (E870s) and adverse reactions to drugs in therapeutic use (E930–E949) were often identified as possible injuries because of the presence of other injury diagnoses. These were dropped in the second stage of the analysis because the injury diagnoses corresponded to these non-injury E codes. Thus, the injury diagnoses were the result, rather than the cause, of the admission.

Invalid diagnosis codes and E codes occurred in the data of nearly every state, with the exceptions of California and Nebraska. Invalid codes ranged from isolated typographical errors to systematic patterns of miscoding. In many cases, we corrected the invalid code by inferring the intended code from other information in the record. Common typos included typing the wrong letter for the first digit of an E code or a V code, switching an 8 for a 9 (or vice versa) in the first digit of a diagnosis or the second digit of an E code, and typing an explicit decimal point after the third digit of a diagnosis (standard coding procedure is to leave the decimal point implicit rather than explicit).

Some codes were corrupted when one hospital attempted to correct or supplement its records using an automated search-and-replace routine. Several hospitals from one state applied a default E code to any record that did not have a legitimate E code. In some cases, this was easy to detect. For instance, some hospitals used E800.0 (railway accident involving collision with rolling

stock: railroad employee) or E999 (late effect of injury due to war operations) as their default E codes. High concentrations of these rarely used codes in a single hospital drew suspicion that they were misapplied.

Another frequent source of invalid diagnoses was the omission of the final digit of a code. ICD-9-CM coding standards specify whether a given code is to have three, four, or five digits, and it must have all digits to be valid. Sometimes when coders lack sufficient information to fill in all digits, they leave the last digit blank instead of coding the digit to indicate an “unspecified” diagnosis or cause. In other cases, it appears that the coder used software that treated the codes as numeric data, rather than character data, and omitted trailing or leading zeroes, or both. Finally, in poisoning cases, we often found inconsistency between diagnosis codes and E codes.

Though coding standards are suggested by Medicare (the Medicare Uniform Billing [UB-92] standards), not all states follow these guidelines. Standardized coding recommendations exist for most HDD fields useful in epidemiological analysis (*Injury Surveillance Workgroup, 2003*), but some states ignored them. *Table 4* summarizes codebook deficiencies in the 19 states in this study, plus 9 additional jurisdictions with readily accessed codebooks—Colorado, Connecticut, the District of Columbia, Iowa, Kentucky, Minnesota, Missouri, North

Table 4
Provision of standard hospital admission and discharge variables among hospital discharge data systems in 25 states and the District of Columbia

State	Diagnosis fields	E-code fields	Procedure fields	ZIP code	Hospital ID	Admit type	Admit source	Discharge status	Race/Hisp
AZ ^a	9	2	6	Yes	Yes	Standard	Standard	Standard	Yes
CA ^a	25	5	21	3-digit	Yes	Non-std	Non-std	Non-std	Yes
CO	15	0	15	Yes	Yes	Standard	Standard	Standard	Yes ^b
CT	10	0	10	Yes	Yes	No	Standard	Standard	Yes
DC	24	0	24	Yes	Yes	Standard	Standard	Standard	Yes ^c
FL ^a	10	0	10	Yes	Yes	Standard	Standard	Standard	Yes
KY	9	2	6	Yes	Yes	Standard	Standard	Standard	No
MA ^a	9	1	10	Yes	Yes	Standard	Standard	Standard	Yes
MD ^a	15	1	15	Yes	Yes	Non-std	Non-std	Non-std	Yes
ME ^a	10	0	10	No	Yes	No	No	Standard	No
MI ^a	7	0	7	Yes	No	Standard	Standard	Standard	Yes ^b
MO	9	1	6	Yes	No	Non-std	Non-std	Standard	Yes
NC	10	0	10	Yes	Yes	Standard	Standard	Standard	Yes ^b
NE ^a	9	1	0	No	No	No	No	Standard	No
NH ^a	10	1	6	No	Yes	Standard	Standard	Standard	Yes
NJ ^a	10	1	8	Yes	Yes	Standard	Standard	Standard	Yes
NY ^a	15	2 ^d	15	Yes	Yes	Standard	Standard	Standard	Yes
OK	16	1	16	Yes	Yes	Standard	Standard	Standard	Yes
PA ^a	9	1	6	Yes	Yes	Standard	Standard	Standard	Yes
RI ^a	7	0	10	Yes	Yes	Standard	Standard	Standard	Yes
SC ^a	10	2 ^d	10	No	No	Standard	Standard	Standard	Yes
UT ^a	9	1	6	Yes	Yes	Standard	Standard	Standard	No ^e
VA ^a	9	1	6	Yes	Yes	Standard	Standard	Standard	Yes ^b
VT ^a	10	0	10	Yes	Yes	Non-std	Non-std	Non-std	Yes
WA ^a	9	1 ^f	6	Yes	Yes	Standard	Standard	Standard	No
WI ^a	9	1	6	Yes	Yes	Standard	Standard	Standard	No

^a Analyzed for this study.

^b Often missing (13–22% of records).

^c Hispanic ethnicity is not coded.

^d The second E code is reserved for location (E849).

^e Usually missing (78%), even though a field is provided.

^f Washington does not put secondary E codes in diagnosis fields.

Carolina, and Oklahoma. (We have cleaned at least 1 year of data from all of these states except CT, IA, and MN.)

In addition to the 19 states analyzed for this paper we have looked at hospital discharge data code books for 6 states and the District of Columbia and have made a limited assessment of data quality. Table 4 describes the provision of standard admission and discharge codes among the hospital discharge data in these 25 states and the District of Columbia. Admission type (urgent or elective) helps analysts to count unique injury events, but it was unavailable in three states and in a non-standard format in 4. Admission source, which identifies transfers and also is important in counting unique injury events, was unavailable in two states and in a non-standard format in 4. Three states did not use the standard discharge codes, which forced recoding to create a merged file that identified deaths in hospital and discharges to nursing home. No standardized or recommended codes exist for source of payment. Consequently payer codes varied widely between states, although all captured payer information.

Non-standard coding made inter-state combination and comparison of data problematic. For example, most states reported admission type, admission source, and discharge status (patient disposition) following the common UB-92 standards, but a few did not (Table 4). It is becoming standard to report race and Hispanic ethnicity as separate variables, though some states still combine them in a single variable. Partially due to privacy concerns, some small states did not report race at all, others reported race in a truncated way (e.g., white versus non-white), several states provided fields to code race/ethnicity but rarely used them, and one state did not provide a code for Hispanic ethnicity. Most states (16) reported a hospital identifier.

4. Discussion

Our findings emphasize that capturing all injury-related cases requires scanning secondary diagnosis fields. Only three-quarters of the injury-related cases identified in the 19-state hospital discharge dataset were coded with a traditional injury diagnosis (ICD 800–994) in the primary diagnosis field. However, by scanning two additional diagnosis fields and all E-code fields, we identified more than 247,000 additional probable injury cases. The additional cases increase the total injury count by 28%. This suggests that the STIPDA recommendation to limit identification of an injury to cases with an injury diagnosis in the primary diagnosis field is too narrow and substantially underestimates the number of injury cases.

State hospital discharge datasets contain information valuable to injury control researchers and practitioners. It is possible to combine multiple states' data if researchers are aware of some of the challenges they may encounter.

E coding in the 19-state file was fairly complete with, overall, 88.1% of probable acute injury cases E coded. However, several states achieved high levels of E coding by using non-specific E codes, which are less useful to researchers. Other important challenges encountered were E coding of cases where no injury occurred (particularly for falls and overexertion) and miscoding of the E code.

The finding that falls and overexertion E codes are the most commonly misused E codes in non-injury cases is consistent with other studies. In principle, a fall should be E coded only if it causes an injury that is medically treated. If a patient falls down *as a result of* an illness or poisoning, but does not sustain an injury from the fall, then the fall should not be coded in the patient's record. But we often found records E coded as falls where the only diagnoses are heart conditions. Overexertion is intended to be used to indicate a cause of sprains, strains, and musculoskeletal conditions. But, like falls, it often accompanied episodes of heart disease.

An interesting finding was that some states attained their high E-coding rates by relying heavily on non-specific E codes, such as E928.9 (unspecified accident). These codes are generally not useful to researchers because they lack detail. Langlois et al. (1995) using Rhode Island hospital discharge data, found that the medical record gave enough information to assign a specific E code to 70% of injuries for which no E codes or vague (non-specific) E codes were recorded. A more recent study of Washington state hospital discharge data found that approximately 20% of injuries assigned to non-specific categories could have been coded to a specific mechanism (LeMier et al., 2001). Record coders, therefore, may be able to make better use of existing documentation to improve E coding.

The inconsistency we found between diagnosis codes and E codes in poisoning cases is a preventable problem. To a great extent, poisoning diagnoses and unintentional poisoning E codes are redundant in the ICD-9-CM codebook. In principle, the same substances should be coded in both fields of the record. In reality, one must capture the information in both fields to get a complete exposure description.

We found that the hospital identifier was particularly important for cleaning hospital data. Many coding errors followed facility-specific patterns. Hospital identifiers allowed us to narrow down the cases that we needed to examine in detail to try to correct the problem.

4.1. Recommendations to data administrators

States that want to improve the usefulness of their hospital discharge data can do several things. Some of these involve data design or quality control at the state level, while others consist of direction to hospitals and coders.

1. Consensus standards are needed for codes for type of payer and race/ethnicity in order to increase code compatibility. These codes need to collapse as necessary to assure confidentiality.
2. Provide hospital identifiers, which are essential for tracking down facility-specific patterns of coding errors. Also distinguish between types of hospitals or types of care—e.g., acute, rehabilitation, psychiatric, long-term. Some states limit their data to acute-care hospitals, while many include other types of hospitals as well. Identifying the type of hospital or type of care would facilitate consistent inter-state comparisons.
3. Differentiate between first admissions and readmissions. This would prevent double counting of injuries. In

combination with an encrypted patient or case identifier, it would also allow examination of the full extent of treatment for a given injury episode.

4. Include dedicated data fields for at least two E codes, and preferably three—a primary, a secondary, and a location. Dedicated fields are necessary because when states list only 9 or 10 diagnosis fields, as is typical, they often fill these fields with diagnosis codes, leaving no room for E codes.
5. Give clear, consistent direction to hospitals and coders about requirements and expectations, and enforce compliance with these requirements.
6. Provide clear documentation to both coders and users of data.
7. Perform quality checks on data as they are received from the hospitals. If too many errors are found, not enough E codes, or excessive reliance on non-specific codes, send the data back for correction. Also check for duplicate records, alert the relevant hospitals to the problem, and eliminate the duplicates.
8. When providing data to researchers, do not limit the dataset arbitrarily—e.g., do not limit the dataset to state residents or to cases based on a generic definition of “injury.”

4.2. Recommendations to data users

1. Do not rely solely on the primary diagnosis for identifying injuries. Scan three diagnoses for an injury diagnosis and all diagnosis fields for an injury E code, or you will miss more than 20% of injury-related hospital admissions.
2. Always sort E codes into separate fields from diagnoses. This will typically require four or five E-code fields, plus a location (E849) field.
3. Process ICD codes – diagnoses, V codes, E codes, and procedures – as character data, rather than numeric. Treating them as numeric data can cause leading and trailing zeroes to be dropped.
4. Develop automated validity checks for ICD codes. Do not automatically delete records with invalid codes. Such deletions could introduce bias into the dataset.
5. Look for duplicate records and drop them.
6. Upon noticing a pattern of errors in the data, always check to see if it is facility-specific. Also check to see if the errors are limited to a particular time period.
7. Process the data in three steps: (1) the full hospital dataset, without data corrections; (2) a broadly defined injury dataset,

whose cases meet any of the criteria for identification as an injury; and (3) a narrowly defined injury dataset, with all non-injuries removed. In our experience, the final, narrow dataset must often be re-created as data errors are detected and the selection process is refined. It is easier to start from the broad injury dataset than to go all the way back to the full hospital dataset.

Acknowledgments

This project was funded in part by grant 1 R01 MH60622 from the National Institute of Mental Health and by grant 1998-WT-UX-0016 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

References

- Guyer, B., Berenholz, G., Gallagher, S.S., 1990. Injury surveillance using hospital discharge abstracts coded by external cause of injury (E code). *J. Trauma* 30 (4), 470–473.
- Injury Surveillance Workgroup, 2003. Consensus recommendations for using hospital discharge data for injury surveillance, online ed. The State and Territorial Injury Prevention Directors' Association, Marietta, GA.
- Langlois, J.A., Buechner, J.S., O'Connor, E.A., Nacar, E.Q., Smith, G.S., 1995. Improving the E coding of hospitalizations for injury: do hospital records contain adequate documentation? *Am. J. Public Health* 85 (9), 1261–1265.
- LeMier, M., Cummings, P., West, T.A., 2001. Accuracy of external cause of injury codes reported in Washington State hospital discharge records. *Inj. Prev.* 7 (4), 334–338.
- Marganitt, B., MacKenzie, E.J., Smith, G.S., Damiano, A.M., 1990. Coding external causes of injury (E codes) in Maryland hospital discharges 1979–1988: a statewide study to explore the uncoded population. *Am. J. Public Health* 80 (12), 1463–1466.
- Mayer, T., Walker, M.L., Johnson, D.G., Matlak, M.E., 1981. Causes of morbidity and mortality in severe pediatric trauma. *J. Am. Med. Assoc.* 245 (7), 719–721.
- Muelleman, R.L., Hansen, K., Sears, W., 1993. Decoding the E code. *Nebr. Med. J.* 78 (7), 184–185.
- Smith, S.M., Colwell, L.S.J., Sniezek, J.E., 1990. An evaluation of external cause-of-injury codes using hospital records from the Indian Health Service, 1985. *Am. J. Public Health* 80 (3), 279–281.
- Sniezek, J.E., Finklea, J.F., Graitcer, P.L., 1989. Injury coding and hospital discharge data. *J. Am. Med. Assoc.* 262 (16), 2270–2272.
- U.S. Department of Health and Human Services (DHHS), 1991. The International Classification of Diseases, ninth revision, Clinical Modification, ICD-9-CM, fourth ed. U.S. Department of Health and Human Services, Washington, DC.