Discovering Knowledge on Pediatric Fluid Therapy and Dysnatremias From Quantitative Data Found in Electronic Medical Records

Steve L. Pham, MD1, Jonathan P. Bickel, MD1,2,4, Michael L. Moritz, MD2,4, James E. Levin, MD, PhD1,2,4
1Department of Biomedical Informatics, 2Department of Pediatrics, 3University of Pittsburgh School of Medicine, Pittsburgh, PA; 4Children’s Hospital of Pittsburgh of UPMC, Pittsburgh, PA

Abstract

It is accepted that intravenous fluid (IVF) therapy can result in hospital-acquired dysnatremias in pediatric patients, with associated morbidity and mortality. There is interest in improving IVF therapy to prevent dysnatremias, but the optimal approach is controversial. In this study, we develop Natremia Deviation and Intravenous Renderer (NaDIR), a tool that preprocesses large volumes of electronic medical record data obtained from an academic pediatric hospital in order to analyze (1) IVF therapy, (2) the epidemiology of dysnatremias, and (3) the impact of IVFs on changes in serum sodium ($\Delta S_{Na}$). We then applied NaDIR to 3,256 inpatient records over a 3 month period, which revealed (1) a 19.9% incidence of dysnatremias, (2) a significant increase in lengths of stay associated with dysnatremias, and (3) a novel linear relationship between $\Delta S_{Na}$ and IVF tonicity. This demonstrates that EMR data that can be readily analyzed to discover epidemiologic and predictive knowledge.

Keywords: Knowledge Discovery, Preprocessing, Hyponatremia, Hypernatremia, Fluid Therapy

Introduction

Intravenous fluid (IVF) therapy is a common practice in pediatric inpatient populations. It is accepted that IVF management is a major cause of serious serum sodium ($S_{Na}$) derangements.1,2 Both dysnatremias are associated with significant neurologic morbidities.3 There is emerging data to suggest that dysnatremias are associated with an increased length of stay (LOS) and cost in those admitted for other reasons.4 While IVF management guidelines have been proposed for preventing dysnatremias,5 how to optimally correct $S_{Na}$ remains controversial.6 Current recommendations include adjusting IVF therapy based on regularly sampled $S_{Na}$.7 Conceptual formulas for modeling changes in $S_{Na}$ ($\Delta S_{Na}$) have been proposed to optimize IVF therapy, but none have been validated on individual patient cases. These models require variables, such as exchangeable sodium and potassium, that are difficult to measure and may not behave in the clinical realm as they do in theory. While attempts to take conceptual $S_{Na}$ formulas to the clinical world have not been successful, we should not be deterred from taking the clinical world to formulas. Knowledge discovery is a process that starts with empirical observations and infers unbiased relationships from them. These observations, or raw data, are made suitable for analysis by a step, often acknowledged as the most important step in knowledge discovery, called preprocessing.8 It employs techniques such as transformation, feature extraction, and classification. Preprocessed data must be validated, sometimes through visual and statistical techniques. This data can then be used to abstract relationships that can be applied to a model that can be clinically tested and validated.

For this study, we hypothesized that knowledge regarding the effect of IVF therapy on dysnatremias in pediatric populations can be discovered from large raw datasets of quantitative information available in electronic medical records (EMRs). We started by extracting retrospective EMR data pertaining to discrete IVF and sodium-containing medication administrations and $S_{Na}$ laboratory values for a cohort of 3,256 inpatient children at an academic pediatric medical institution. We then developed Natremia Deviation and Intravenous Renderer (NaDIR), a preprocessing tool that transformed raw quantitative data vectors into data objects and analyzed these data objects for the characteristics of IVF therapies and demographics of patients with dysnatremias in an unbiased fashion. NaDIR outputted datasets for further processing and visual plots for validation of aggregate and individual data. We then processed this output and found (1) dysnatremia incidence and demographic data, (2) an increase in LOS associated with instances of dysnatremias, (3) a significant increase in $S_{Na}$ after 48 ±12 hours of hypertonic IVFs, and (4) a novel linear relationship between $\Delta S_{Na}$ and cumulative IVF sodium concentration ($\Sigma IV_{Na}$).

Methods

Data: Datasets were obtained from the clinical data warehouse at Children’s Hospital of Pittsburgh of UPMC, a large academic pediatric medical institution. The EMR (Cerner Millennium®) stores
records for patient demographics, encounter characteristics, medication administrations, and laboratory values in a relational database. De-identified data were extracted as a Comma Separated Value (CSV) file using a structured SAP BusinessObjects query:

1. Data: Study Number, Admit Age (in days), Clinical Minutes From Admission, Managing Service, Clinical Event (names of medication administrations and laboratory values), Clinical Event Result (quantitative values), Clinical Event Result Units (measurement units)

2. Predicates: Admission Date within a time range, Discharge Date before a specified date, Clinical Events matching specified medications (IVFs and sodium-containing medications diluted into IVFs) and laboratory values (serum electrolytes including $S_{Na}$)

3. Ordered By: Study Number, Admit Age, and Clinical Minutes From Admission

Medication predicates were established by surveying EMR medication identifiers. Medication types included IVFs as well as medications containing “sodium” or “Na” in their names. Temporal data was represented as “Clinical Minutes From Admission” to remove identifiable dates.

**Software Design:** NaDIR was developed using Visual C# in a .NET environment. The input requirement was the CSV file of the data warehouse extract. The software user could specify output modalities that included (1) images of plots for validation, (2) images of histograms, and (3) preprocessed CSV files for knowledge discovery.

**Data Model:** NaDIR transformed clinical event vector data into data objects, which were then classified, transformed, and used for feature extraction. Data objects were created for patient records, discrete temporal entities, clinical events, and medication data. Figure 1 illustrates the structures of these objects.

**Feature Extraction:** NaDIR’s feature extraction modalities included: (1) calibration of IVF delivery rates by specified time intervals such as 24H and by estimated body surface area, (2) calculation of cumulative IVF sodium mEq content, cumulative IVF volume, and dividing the two to calculate $\Sigma IV_{Na}$, (3) all possible paired vectors of $S_{Na}$ before ($S_{Na(pre-IV)}$) and after ($S_{Na(post-IV)}$) IVF interventions, and (4) incidence of each medication stratified by hour of initiation, delivery rates, and durations of therapy.

**Output:** NaDIR’s output modalities included (1) image files plotting each patient’s cumulative IVF sodium mEq content, cumulative IVF volume, and $S_{Na}$ time series with medications along a timeline, (2) image files of histograms of each medication incidence stratified by delivery rates and durations, (3) a CSV file containing a matrix of laboratory values, medication administrations, and feature extractions indexed by a user-specified time interval, (4) a CSV file of age and LOS demographics and incidences of medication administrations stratified by volumes, rates, and durations, and (5) a CSV file of all possible $S_{Na(pre-IV)}$ and $S_{Na(post-IV)}$ paired vectors for all patients with the intervening IVF therapy’s duration and $\Sigma IV_{Na}$.

In practice, IVFs are delivered continuously. However, nurses record IVF administrations as discrete volume values at time intervals, usually every 60 minutes. So while the study patient and temporal data structures were mapped directly from values in the raw data vectors, uninterrupted medication data could not be. Discrete clinical event time series were transformed into continuous medication data objects using this algorithm:

\[
\text{iterate through the medication calendar} \\
\text{if a clinical event matches an existing medication ending 60 minutes ago then} \\
\text{extend the existing medication end time by 60 minutes} \\
\text{else} \\
\text{instantiate a new medication object}
\]

**Classification:** NaDIR classified patient subsets within the dataset. It identified (1) hyponatremic patients as those with any episode of $S_{Na} < 135$ mEq/L, (2) hypernatremic patients with any episode of $S_{Na} > 145$ mEq/L, and (3) non-dysnatremic patients without episodes of (1) or (2) and having $135 \leq S_{Na} \leq 145$ mEq/L for the entirety of the hospital stay. A patient with multiple hyponatremic or hypernatremic $S_{Na}$ data entries would have each time point classified as a separate episode. A patient can be classified in both hyponatremic and hypernatremic subsets if the patient had temporally separate episodes of each.
**Analysis:** We then analyzed patient demographics and $S_{Na}$ paired vectors with R. We used one-way Analysis of Variance (ANOVA) to validate differences in LOS and ages between hyponatremic, hypernatremic, and non-dysnatremic patients.

We used $S_{Na}$ pairs to observe changes in $S_{Na}$ in response to intervening IVFs. We restricted $S_{Na}$ pairs to a temporal interval of 48 ±12 hours. Because serum electrolytes are typically drawn at 24 hour intervals, we chose 48 hours to reflect more than one day of continuous IVF. We chose a ±12 hour range because $S_{Na}$ values are not always obtained every 24 hours. We then classified $S_{Na}$ pairs into three subsets by $\Sigma IV_{Na}$: (1) hypotonic when $\Sigma IV_{Na} < 140 \text{ mEq/L}$, (2) isotonic when $140 \text{ mEq/L} \leq \Sigma IV_{Na} \leq 155 \text{ mEq/L}$, and (3) hypertonic when $\Sigma IV_{Na} > 155 \text{ mEq/L}$. We based this classification on the clinical use of normal saline (154 mEq/L) and Plasma-Lyte (140 mEq/L) as isotonic solutions. We allow the upper-limit to be 155 instead of 154 because concurrent medications, such as sodium tabs, caused small concentration effects.

Validation was performed on both the whole set of $S_{Na}$ pairs and between the tonicity classified subsets. We calculated $\Delta S_{Na}$ from the $S_{Na}$ pairs and used a Shapiro-Wilk test to assess the distribution of $\Delta S_{Na}$ for normality. Then, we used paired Student’s t-tests to validate whether the $S_{Na(pre-IV)}$ and $S_{Na(post-IV)}$ values in each subset did have a significant differences before and after 48 ±12 hours of intervening IVF.

To study relationships between $\Delta S_{Na}$ and $\Sigma IV_{Na}$, we performed a linear regression analysis with the assumption that a normally distributed set of $\Delta S_{Na}$ might depend proportionally on $\Sigma IV_{Na}$.

**Results**

NaDIR’s generated plots for visual validation of each patient. In figure 2 is one patient’s visual history of cumulative IVF sodium mEq content and volume overlapped with a $S_{Na}$ time series. The bars represent continuous IVFs. Normal saline solutions coincided with increases in serum sodium, while hypotonic solutions coincided with decreases in serum sodium.

The data warehouse query produced a set of 926,448 medication and laboratory entries pertaining to 3,256 patients discharged over 3 months.

Table 1 presents incidence and demographic data. Incidence of hyponatremia was 17.1%, while incidence of hyponatremia was 5.1%. Incidence of any dysnatremia was 19.9%. Patients with hypernatremia were on average younger than patients with hyponatremia and patients without dysnatremia by 1 year, with ANOVA (p < 0.05). Patients with hyponatremia, hypernatremia, and no dysnatremia averaged LOS’s of approximately 13, 22, and 4 days, respectively. Differences in length of stay were confirmed by ANOVA (p < 10^-6). The Pediatric ICU was the most frequent site of dysnatremic episodes.

---

**Figure 2:** A plot of a patient’s $S_{Na}$, maintenance IVF sodium content, and maintenance IVF volumes with medication bars.
Validity testing using the Shapiro-Wilk test of all $\Delta S_{Na}$ showed a normal distribution ($W = 0.974$, $p < 10^{-6}$). In Table 2, the paired Student’s t-tests demonstrated significant differences between $S_{Na(pre-IV)}$ and $S_{Na(post-IV)}$ at 48 ±12 hours classified by concurrent cumulative IV tonicities of hypertonic, isotonic, and hypotonic.$\Delta$SNa is in units of mEq and $\Sigma$IVNa is in units of mEq/L. The correlation coefficient’s standard error (SE) was 2.06 $\times 10^{-3}$ and $p = 6.66 \times 10^{-2}$,

$$\Delta S_{Na} = 4.94 \times 10^{-3} \times \Sigma IV_{Na} - 4.05 \times 10^{-1}$$

where $\Delta S_{Na}$ is in units of mEq and $\Sigma IV_{Na}$ is in units of mEq/L. The correlation coefficient’s standard error (SE) was 2.06 $\times 10^{-3}$ and $p < 0.05$, and the y-intercept’s SE was 2.20 $\times 10^{1}$ and $p = 6.66 \times 10^{-2}$.

### Discussion

Our knowledge discovery process provided numerous insights. First, several factors demonstrated the validity of our sample. Plots of $S_{Na}$ showed trends in response to IVF medications. As is commonly experienced with visualization modalities, our graphs reduced the cognitive overload of processing complex and disparate data types. With hyponatremia defined as $S_{Na} < 135$ mEq/L, we found a 17.1% incidence at our institution in concordance with existing literature using a similar definition of hyponatremia.$^{10}$ With hypernatremia defined as $S_{Na} > 145$ mEq/L, we found a 5.1% incidence at our institution. One study of hypernatremia in a pathology chemistry laboratory database at a pediatric institution used a more restrictive cutoff of 150 mEq/L and found a smaller incidence of 1.4%.$^{1}$

Demographically, the finding that the pediatric ICU was the most common site of dysnatremias agrees with previous studies.$^{7}$ The findings that those with hypernatremia were receiving coincident hypertonic IVFs and that those with hyponatremia were receiving coincident hypotonic IVFs agrees with our physiologic understanding of IVF tonicity.

Demographic data also illustrated new findings. Hypernatremic patients averaged a year of age less than hyponatremic and non-dysnatremic patients. The increased LOS’s associated with dysnatremic patients is also significant. Two studies of heart failure patients have also shown an increased LOS with hyponatremia.$^{4}$ That those with hypernatremia and hyponatremia stayed 18 and 9 days longer, respectively, than non-dysnatremic patients is new data and warrants further investigation. A longer LOS coincident with a dysnatremic episode may be due to other coincident disorders.

Our analysis of NaDIR’s $S_{Na(pre-IV)}$ and $S_{Na(post-IV)}$ paired vectors revealed significant relationships in an unbiased fashion. Paired t-tests established an association between hypertonic IVF and positive $\Delta S_{Na}$ at 48 ±12 hours. The subset size is small, possibly indicating that the sustained use of hypertonic IVFs for durations of 48 ±12 hours is infrequent at our institution. The linear regression analysis produced a statistically significant correlation coefficient despite using only one variable of $\Sigma IV_{Na}$. The y-intercept, however, was not significant. Using this equation, predictive ranges of $S_{Na}$ after 48 ±12 hours of sustained IVF administration can be calculated if the cumulative IVF sodium and volume contents are known.

These findings demonstrate the ability of knowledge discovery to thoroughly analyze large sets of quantitative EMR raw data, validate previous findings, and generate new relationships.

### Table 1: Demographics of subsets classified by those who had at least one episode of hyponatremia or hypernatremia, and no dysnatremias. Age of Admission, Length of Stay, and Corresponding IVF tonicity are presented as means.

<table>
<thead>
<tr>
<th>Study Groups</th>
<th>Number</th>
<th>Age of Admission (Years)</th>
<th>Length of Stay (Days)</th>
<th>Most Frequent Managing Service During Episodes</th>
<th>Episodes While on IVFs</th>
<th>Corresponding IVF Tonicity (mEq/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyponatremic</td>
<td>560</td>
<td>7.44</td>
<td>13.22</td>
<td>3256</td>
<td>1123</td>
<td>73</td>
</tr>
<tr>
<td>Hypernatremic</td>
<td>166</td>
<td>6.44</td>
<td>21.71</td>
<td>3256</td>
<td>1067</td>
<td>181</td>
</tr>
<tr>
<td>No dysnatremia</td>
<td>2607</td>
<td>7.56</td>
<td>4.15</td>
<td>3256</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>All</td>
<td>3256</td>
<td>7.52</td>
<td>5.90</td>
<td>3256</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

ANOVA F-statistic: $\uparrow p < 0.05$  $\uparrow\uparrow\uparrow p < 10^{-6}$

### Table 2: Paired t-tests validating differences between $S_{Na(pre-IV)}$ and $S_{Na(post-IV)}$ at 48 ±12 hours classified by concurrent cumulative IV tonicities of hypertonic, isotonic, and hypotonic.

<table>
<thead>
<tr>
<th>Fluid Tonicity</th>
<th>Number of Pairs</th>
<th>$\Delta S_{Na(post-pre)}$ (mEq/L)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypotonic</td>
<td>595</td>
<td>-0.198</td>
<td>2.50x10^{-1}</td>
</tr>
<tr>
<td>Isotonic</td>
<td>181</td>
<td>0.492</td>
<td>8.49x10^{-2}</td>
</tr>
<tr>
<td>Hypertonic</td>
<td>12</td>
<td>1.917</td>
<td>1.53x10^{-2}</td>
</tr>
</tbody>
</table>

$\uparrow p < 0.05$
There are limitations to our process. A common pitfall in knowledge discovery is the validity of the source data. Laboratory data extracted from EMRs depend on the accuracy of data entry within the health system. Our query extracts those patients admitted and discharged within 3 months. Consequently, our patient dataset may contain a seasonal bias.

NaDIR’s preprocessing can also improve. Some enhancements include identifying and scrubbing out inaccurate, incomplete, or outlier records. NaDIR does not preprocess some known factors: serum potassium, urine electrolytes, clinical signs such as neurologic sequela of dysnatremia, disorders that affect serum sodium such as SIADH and acute kidney injury, and managing services known to have a high incidence of dysnatremias such as the pediatric ICU. NaDIR currently lacks the ability to extract these features. Natural language processing can extract nominal data such as incidences of neurologic sequelae and SIADH.

We believe that the knowledge discovery process represents a new paradigm in S Na modeling. Whereas most existing models start with concepts, we demonstrate a process that uses raw data to empirically generate a significant model. Our linear regression analysis is only a first attempt at modeling S Na. As a single variable correlation, it is relatively simplistic. If NaDIR can extract more feature variables, they can be included into a multivariate regression model. A time-series analysis on S Na may introduce trend prediction into the model.

With continual refinement, we believe that we can discover further knowledge that will improve our model. Validation of this evolving model will need to be performed on other retrospective datasets and in prospective trials. A comprehensive model of S Na may eventually be capable of real-time detection of iatrogenic dysnatremias. Such a model could suggest changes in practices to help optimize IVF therapy, reduce dysnatremia incidences, and improve pediatric inpatient outcomes.

Author Contributions

SP coded NaDIR and performed R analysis. JB and JL acted as mentors for SP. JL extracted all datasets, MM acted as a clinical domain expert. SP performed a literature review and wrote the manuscript. All authors reviewed and approved the manuscript.

Acknowledgements

The authors would like to thank Dr. Atul Butte for his guidance, and Dr. Chris Longhurst and Dr. Scott Sutherland for their support. Study PRO09060168 was determined to be exempt by The Institutional Review Board of the University of Pittsburgh Medical Center. The Children’s Hospital of Pittsburgh provided operational resources.

Citations