Statistical Analysis Plan H9X-MC-B021

Crossover Study Comparing Dulaglutide (Trulicity) Pen and the Semaglutide (Ozempic) Pen

NCT03724981

Approval Date: 09-Oct-2018

1. Statistical Analysis Plan:
H9X-MC-B021: Crossover Study Comparing the Dulaglutide (Trulicity) Pen and the Semaglutide (Ozempic) Pen

dulaglutide (LY2189265) T2DM

Study H9X-MC-B021 is an open-label, multicenter, randomized, crossover study assessing patient preference (and training time) for the dulaglutide pen versus the semaglutide pen among a sample of self-injection naïve patients with type 2 diabetes mellitus who are taking oral medication.

Eli Lilly and Company
Indianapolis, Indiana USA 46285

Protocol H9X-MC-B021

Statistical Analysis Plan electronically signed and approved by Lilly on date provided below.

Approval Date: 09-Oct-2018 GMT

2. Table of Contents

Section **Page**

3. Revision History

Statistical Analysis Plan (SAP) Version 1 was approved prior to the first visit when a subject receives study drug or any other protocol intervention.

4. Study Objectives

4.1. Primary Objective

The primary objective is to determine patient preference between the dulaglutide and semaglutide injection devices.

4.2. Secondary Objectives

A gated secondary objective of this study is to compare the two injection devices with regard to ease of use. This objective will be tested if a significant difference is found in the primary objective.

4.3. Exploratory Objectives

This study has four exploratory objectives: (1) to collect data to support the preference question in the primary objective; (2) to compare time to train on the dulaglutide and semaglutide devices; (3) to examine patients' willingness to use each of the devices after being trained on both, and; (4) to validate the Diabetes Injection Device Preference Questionnaire (DID-PQ).

5. Study Population and Determination of Sample Size

5.1. Selection of Study Population

5.1.1. Inclusion Criteria

Patients are eligible to be included in the study if they meet all the following criteria at screening:

Participant Characteristics

[1] Are at least aged ≥18 years at the time of screening

[2] Diagnosed with type 2 diabetes

[3] Self-injection naïve to all injectable treatment (for example, diabetes therapies and other medical conditions)

[4] Injection naïve to performing all injectable treatment (for example, diabetes therapies and other medical conditions) to others

[5] Must **currently receive oral** treatment for their type 2 diabetes

[5a] **Pilot phase participants only:** must be able to **bring proof of their oral treatment prescription for type 2 diabetes to the interview** (e.g., the medication itself, the medication packaging, a prescription note, or a letter from their doctor). This criterion does not apply to the main phase because main phase participants will have diagnoses confirmed by the clinical sites with awareness of the patients' diabetes and treatment.

[6] Willing and able to attend an in-person interview session

[7] Able to read, speak, write, and understand the English language

[8] Able and willing to give signed informed consent prior to study entry

[9] Able to complete the protocol requirements

5.1.2. Exclusion Criteria

Participants will be excluded from study enrollment if they meet any of the following criteria at screening:

Medical Conditions

[1] Currently diagnosed with gestational diabetes and/or type 1 diabetes

[2] Cognitive or physical difficulties that could interfere with ability to understand the training, perform the injection tasks, or complete the study questionnaires as judged by the investigator

Prior/Concurrent Clinical Trial Experience

[3] Are currently enrolled in any other clinical study involving an investigational product or any other type of medical research judged not to be scientifically or medically compatible with this study

[4] Have participated, within the last 30 days, in a clinical study involving an investigational product.

Other Exclusions

[5] Is a health care practitioner who is trained in giving injections

[6] **For Main Phase participants only:** Investigator, site personnel or immediate family member of investigator or site personnel. Immediate family is defined as a spouse, parent, child, or sibling, whether biological or legally adopted

[7] Is an employee of any of the following companies: Eli Lilly and Company, Novo Nordisk, Evidera or PPD

[8] Currently pregnant

5.2. Sample Size Determination

To test the null hypothesis that patients have equal preference for either dulaglutide or semaglutide pen (50% of patients prefer dulaglutide pen and 50% prefer semaglutide pen), a sample size of 260 patients will provide 90% power against alternative hypothesis that there is a preference with 60% of the patients choosing the dulaglutide pen over the semaglutide pen. Calculation is carried out using 2-sided chi-square test (normal approximation) at 0.05 level of significance. Software used is Nquery + nTerim 4.0. Assuming that approximately 10% will not provide any preference information by choosing the "no preference" option on the Global Preference Item, 290 patients will be needed to have at least 260 patients with preference information. The Prescott test which takes into account the order the device is used as well as the choice of "no preference" provides approximately 90% power against the alternative hypothesis of 54% choosing dulaglutide pen, 10% indicating no preference, and the remaining 36% choosing semaglutide pen. This is based on simulation of 1000 runs and assumes that there is no order effect.

5.3. Populations for Analyses

For purposes of analysis, the following populations are defined:

Population	Description
Randomized	Participants determined to be eligible and then assigned to one of 2 device order groups (50% will be randomized to use the dulaglutide device first, while the other half will use the semaglutide device first). More details about randomization for this study are provided in Section 7.2 of the protocol.
Evaluable	Participants for whom device preference can be evaluated: Randomized participants who are exposed to both devices (i.e., participant was shown both devices via either video or demonstration, regardless of whether they successfully complete the training) and complete the Global Preference Item.
Withdrawn	Participants who withdraw from the study before being exposed to both devices and completing the Global Preference item. Details on study withdrawal are provided in Section 8 of the protocol.

6. A Priori Statistical Methods

6.1. General Statistical Considerations

Statistical analysis of this study will be the responsibility of Evidera. The primary analyses will be conducted on the evaluable patient population (i.e., patients who provide a response to the Global Preference Item). No imputation will be performed for missing data, and the analysis will be performed on observed cases.

The significance level for all statistical tests will be $p < 0.05$. A serial gatekeeping strategy will be used to control for type 1 error for the primary and gated secondary objectives. The gated secondary hypothesis will be tested only if the primary null hypothesis (i.e., that there is no difference in the preference for dulaglutide and semaglutide devices) is rejected.

Any change to the statistical analysis methods described in the SAP will require an amendment ONLY if it changes a principal feature of the SAP. Any other change or addition to the statistical analysis methods described in the SAP, and the justification for making the change, will be described in the clinical study report (CSR).

6.2. Adjustments for Covariates

There are no planned adjustments for covariates for the primary or secondary analyses. For exploratory analyses involving Time-To-Train (TTT) as a continuous dependent variable, covariates such as age and gender will be used.

6.3. Handling of Dropouts or Missing Data

Study team members will perform a thorough review of the participants' questionnaires prior to the completion of each interview to minimize missing data. No imputations will be performed for missing data, and the analysis will be performed on observed cases. If a participant decides to withdraw from the study, data collected to the point of withdrawal may be used in analysis to preserve the integrity of the research project. There will be no lost to follow-up data because there is no follow up period in this study.

6.4. Multicenter Studies

Approximately 14 clinical sites will recruit patients for this study. Descriptive statistics (e.g., frequencies, percentages) for patient demographics will be presented by clinical site as described in Section 6.9.

6.5. Multiple Comparisons/Multiplicity

Gatekeeping strategy will be used to control for type 1 error for the primary and gated secondary objective. There are no other plans to adjust for multiplicity.

6.6. Use of "Efficacy Subset" of Patients

Not applicable.

6.7. Active-Control Studies Intended to Show Equivalence

Not applicable.

6.8. Patient Disposition

A detailed description of the patient disposition will be displayed in a figure. The figure will present information including but not limited to: 1) number of eligible patients; 2) number of enrolled patients; 3) number of patients in final analysis sample; 4) number of patients randomized to dulaglutide as the first device in the administration sequence (also referred to as the "dula-sema" group); and, 5) number of patients randomized to semaglutide as the first device in the administration sequence (also referred to as the "sema-dula" group).

6.9. Patient Characteristics and Patient-Reported Outcome Data

6.9.1. Patient Characteristics

The frequency of patients assigned to each device sequence will be presented by clinical site. Descriptive statistics will be used to summarize the demographic and clinical characteristics of the total analysis sample as well as by the assigned device sequence group (i.e., sema-dula and dula-sema). Descriptive statistics will also be presented by TTT assignment group (i.e., assigned to TTT vs. not assigned to TTT), and by individual clinical sites to ensure there are no systematic differences between the TTT and non-TTT groups. Continuous variables (e.g., age) will be summarized with means, standard deviations, ranges, and minimum and maximum values, while categorical variables (e.g., gender, race) will be presented in terms of frequencies and percentages. Patient-reported characteristics for the device sequence and TTT assignment groups will be compared (i.e., t-tests for continuous variables and chi-square tests for categorical variables) to determine if there are significant group differences.

Site-reported clinical characteristics will be presented descriptively; these statistics will be presented by the total analysis sample as well as by the assigned device sequence group, by TTT assignment group, and by individual clinical sites. Site-reported characteristics for the device sequence and TTT assignment groups will be compared (i.e., t-tests for continuous variables and chi-square tests for categorical variables) to determine if there are significant group differences.

To understand the geographic representation of the sample (total sample and by randomization group), counts and percentages of patients recruited from each US geographic region (i.e., Northeast, Midwest, South, West) will be presented.

6.9.2. Patient-Reported Outcomes

Distributional characteristics (frequencies and percentages of responses) for all patient-reported outcome (PRO) data, including data from the Global Preference Item, DID-PQ, MDDAB-semaglutide, and the MDDAB-dulaglutide will be presented by total sample and by assigned device sequence group. Descriptions of these PRO measures are below.

6.9.2.1. Global Preference Item

All participants will complete the Global Preference Item after being trained and using both devices. Participants will be asked, "Overall, which device do you prefer?" Participants will also be asked to explain why they selected their response for Ozempic, Trulicity, or no preference. This item was developed specifically for use in this device preference multicenter crossover study. To help participants remember the devices more accurately, colored images of the devices have been inserted into the questionnaire.

6.9.2.2. Diabetes Injection Device – Preference Questionnaire

The DID-PQ was designed for the purpose of assessing patient preferences between two non-insulin injection devices (Matza et al. 2018a; Matza et al. 2018b). The DID-PQ is completed only by patients who have used devices for at least two injectable medications. In the current study, patients will be using two devices to inject an injection pad, but will not be injecting themselves.

The DID-PQ has 10 items developed based on qualitative research with patients and validated in a psychometric study (Matza et al. 2018a; Matza et al. 2018b). Items 1 to 7 focus on specific characteristics of injection delivery systems, and these seven items comprise the Device Characteristics subscale. Items 8 to 10 are three global items assessing preference based on overall satisfaction, ease of use, and convenience of the injection devices.

On the DID-PQ, each item is rated on a 5-point scale allowing patients to indicate whether they prefer or strongly prefer one of the devices over the other. For each item, patients may also respond by selecting the "no preference" (i.e., neutral) response option, indicating that they have no preference between the two devices. Importantly, the response options of the DID-PQ are not on an ordinal scale assessing a single dimension. These five response options of the DID-PQ comprise a non-linear scale for which mean scores are not calculated.

Specifically, for this study, the DID-PQ response options will be: Strongly Prefer Ozempic Device, Prefer Ozempic Device, Strongly Prefer Trulicity Device, Prefer Trulicity Device, or have No Preference. To help participants remember the devices more accurately, colored images of the devices have been inserted into the questionnaire.

For individual items of the DID-PQ, data will not be imputed. If a response to one of these items is missing, then it is not possible to derive a score for that item.

6.9.2.3. Four Items from the Medication Delivery Device Assessment (MDDAB)

The original MDDAB was adapted from insulin-specific questionnaires and has been modified for use in both injection-naïve and non-insulin requiring participants (Matfin et al. 2015). Four items selected from the MDDAB (Matfin et al. 2015; Yu et al. 2017) will be asked of the participants about each of the two devices under investigation in this study (Ozempic and Trulicity). These additional questions will be completed in the order the participants were randomized to use the devices. The first 3 questions ask participants how easy or difficult it was to: (1) learn to use the device, (2) to follow the instructions when using the device, and (3) overall, how difficult or easy was the device to use? The last question asks participants to

"Please check the number that best indicates how willing you are to continue using the device" on a scale from 1 ("Definitely Unwilling") to 5 ("Definitely Willing").

6.10. Treatment Compliance

Not applicable.

6.11. Concomitant Therapy

Not applicable.

6.12. Efficacy Analyses

This study focuses on preference between devices. No investigational drug or treatment will be administered during this study. Therefore, no efficacy analyses will be conducted. The primary, secondary, and exploratory analyses are detailed below.

6.12.1. Primary Analysis

The primary analysis will examine whether there is a difference in preference between the semaglutide device and dulaglutide device as indicated by responses to the Global Preference Item (i.e., Overall, which device do you prefer?). The Prescott test will be run to determine whether there is a statistically significant difference in preference between the devices, while controlling for order effects and taking into account the neutral responses. This test is based on a comparison of the difference between the number of patients who prefer the first device and the number of patients who prefer the second device in each crossover group (Prescott 1981; Pictor 2003).

For this analysis, the data are conceptualized as a 2x3 contingency table with rows consisting of the two crossover groups and columns indicating preference for the first device used, no preference, or preference for second device used.

Crossover Group	Prefers First Device	No Preference	Prefers Second Device	Total
Dulaglutide-Semaglutide	a	b	c	Row 1
Semaglutide-Dulaglutide	d	e	f	Row 2
Total	Column 1	Column 2	Column 3	Total study N

The test considers the difference in counts of the first and third columns in each group computed as first device minus second device (a - c and d - f). These represent the difference between patients preferring the first device and patients preferring the second device within each crossover group. The difference between differences in numbers of patients ($\Delta = [a - c] – [d - f]$) represents the preference for one device over the other. A mathematically equivalent way to present this analysis is as $\Delta = [a + f] – [d + c]$. In this equation, the term "a + f" represents all

patients preferring the dulaglutide device, and the term "d + c" is the sum of all patients preferring the semaglutide device across the two crossover groups. A positive Δ would indicate more frequent preference for dulaglutide, while a negative Δ would indicate more frequent preference for semaglutide.

The p-value of the test is based on calculating the probability of finding a Δ by chance alone that is equal to or more extreme than (i.e., greater than a positive Δ or less than a negative Δ) the Δ that is observed (when row and column totals are held constant). The device preference results (e.g., counts, percentages) will be presented by total sample and by assigned randomization group using the contingency structure above.

For patients who express a preference on the Global Preference Item, the self-reported reason for this preference will be presented in a separate table. These qualitative data (i.e., open-ended patient responses on the Global Preference Item CRF) will be summarized and collapsed into categories of reasons (e.g., needle size, injection frequency), and the frequencies and percentages of patients reporting each category will be presented by device.

6.12.2. Gated Secondary Analysis

The secondary analysis will compare the dulaglutide and semaglutide devices with regard to ease of use. Responses to the five-point response scale of the DID-PQ global item 9 (Overall ease of use) will be collapsed into three categories with the "strongly prefer" and "prefer" options being combined for each device. The resulting three categories will include: prefer dulaglutide, no preference, and prefer semaglutide. The Prescott test will then be run to examine if a statistically significant difference in preference between the devices exists, while controlling for order effects.

6.12.3. Exploratory Analyses

6.12.3.1. Exploratory 1. DID-PQ Analyses of Items 1-8 and 10

Exploratory analyses of items 1–8 and 10 of the DID-PQ will follow the same procedures as those for the gated secondary objective described above. The responses for each item will be collapsed to create three categories (prefer dulaglutide, no preference, and prefer semaglutide) and the Prescott test will be run to examine whether there is a statistically significant difference in preference between the devices, while controlling for order effects.

6.12.3.2. Exploratory 2. Time-to-Train Analyses (TTT)

Time-To-Train with each device will be measured in a subset of patients. Of the approximately 14 sites expected to recruit patients for the study, four sites will be selected to participate in the TTT assessment. The TTT assessment will require interview facilities with interview rooms allowing for observation from behind a one-way mirror. Therefore, the TTT sites will be selected based on proximity to suitable facilities, while taking into account geographic diversity and the split between general practice and endocrinology sites.

Time-To-Train will be reported for the total time spent for each device, including the time for the video component as well as the time after subtracting the video component. Average duration

required to train each device will be reported and presented for each device (e.g., dulaglutide and semaglutide) and by administration sequence (i.e., dula-sema and sema-dula).

A linear mixed modeling framework accounting for the crossover design with the appropriate terms for device (either dulaglutide device or semaglutide device), sequence (either dula-sema or sema-dula), and period (either trained first or trained second) will be used to compare TTT between the devices The model will include device, sequence, and period as fixed effects and patient as a random effect. Least square means, 95% confidence interval, and p-values will be calculated from this model. If a carryover (sequence) effect is found to be significant, an analysis based on data from the first period only will also be carried out.

The balance in baseline characteristics of patients randomized to the TTT sequence groups will be assessed and appropriate analysis conducted to explain the results.

Linear model assumptions, such as normality and common variance, will be assessed by examining the residuals. Linear models tend to be robust to violations of these assumptions, but if there are any substantial violations of the assumptions, then alternative approaches such as non-parametric methods may be utilized.

6.12.3.3. Exploratory 3. Supplemental Question Analyses

Supplemental Questions are included in the survey to examine willingness to use each of the devices after being trained on both. The Supplemental Questions developed for this study are based on previous items used in a study by Poon et al. (2018). The first question will be administered to participants before the device trainings and will determine the participants' "willingness to use a diabetes medication that required an injection for each dose." The second and third Supplemental Questions will be administered to the participants after all the trainings, mock injections, and other study measures have been completed. Questions S2 and S3 will ask participants specifically about their willingness to use the Ozempic and Trulicity devices, respectively. Descriptive statistics will be reported for each item (i.e., frequencies, percentages).

6.12.3.4. Exploratory 4. Evaluation of the DID-PQ

A psychometric evaluation of the DID-PQ and assessment of various scoring approaches proposed in the DID scoring guide will be performed. The DID-PQ has been administered in two studies thus far: the original psychometric validation study (Matza et al. 2018b) and the US study comparing the dulaglutide and liraglutide devices (Matza et al. 2018c). In both studies, the DID-PQ was administered to a relatively small subgroup of patients who had used more than one GLP-1 receptor agonist device. Therefore, it has not previously been possible to perform psychometric analyses examining the validity of this PRO instrument nor the proposed scoring approaches. Since it is anticipated that the entire sample in this study will complete the DID-PQ, this study provides a unique opportunity to perform analyses evaluating the performance of this instrument.

Three approaches to analyzing the DID-PQ will be examined, as described in sections A, B, and C below:

A. Categorical descriptive analyses

B. Testing for significant differences between devices on individual items of the DID-PQ
C. Aggregate scoring for the seven device characteristics items of the DID-PQ

A. Construct Validity of the Categorical Approach to Analyzing the DID-PQ

In publications thus far, DID-PQ data has been presented descriptively and categorically without additional quantitative analysis. Results have been presented as descriptive statistics for each individual item (i.e., frequency and percent of patients selecting each response). This descriptive approach to analyzing DID-PQ data will first be examined by comparing DID-PQ responses to responses on the global preference item. These analyses are examining the instrument's construct validity, which is the extent to which an instrument performs as expected relative to an ancillary measure.

First, patients' responses to each of the 10 items of the DID-PQ will be collapsed from five levels to three levels, by combining the "prefer" and "strongly prefer" response options for dulaglutide and semaglutide. These collapsed three-level DID-PQ responses will be compared to responses on the Global Preference Item in a frequency table so that agreement between the two instruments can be assessed. Each item on the DID-PQ will be assessed for concordance as indicated by percent agreement, Gwet's AC_1 statistic (Gwet 2014), and the prevalence-adjusted, bias-adjusted kappa (PABAK) statistic (Byrt et al. 1993).

The AC_1 is an agreement statistic that is similar to kappa, but uses a different definition of chance agreement. The AC_1 is designed to be less affected by high prevalence or marginal imbalance compared to kappa. This approach yields a slightly conservative estimate of agreement, because it assumes the scale is nominal and any disagreement is an equal amount of disagreement. The PABAK statistic takes a slightly different approach to dealing with either high prevalence or bias in marginal distributions. The PABAK calculates its estimate of chance agreement using different assumptions regarding high prevalence or bias in marginal distributions than the plain kappa.

After these 3x3 tables, a series of 5x3 tables (one for each DID-PQ item) will be presented to compare the un-collapsed 5-level categorical responses of the DID-PQ to the Global Preference Item. Because of the unequal number of response options in this comparison, the summary statistics will not be run. Instead, results will provide a general picture of whether all five levels of the DID-PQ are performing as expected. For example, it is expected that patients who choose either prefer dulaglutide or strongly prefer dulaglutide on the DID-PQ would also tend to report a preference for dulaglutide on the global preference item.

Construct validity will further be assessed by comparing DID-PQ responses to responses on four items selected from the MDDAB (Matfin et al. 2015; Yu et al. 2017). The four items of the MDDAB assess (1) ease of learning how to use the device, (2) ease of following instructions when using the device, (3) overall ease-of-use, and (4) willingness to continue using the device. These analyses will be run with selected DID-PQ items hypothesized to be related to these four constructs assessed by the MDDAB.

B. Statistical Significance of Differences in Preference for Each Item of the DID-PQ

After reporting DID-PQ findings descriptively, it may also be useful to determine whether significantly more respondents preferred one device over the other with regard to specific items of the DID-PQ. The current crossover study will be the first sample with a large enough sample size to perform and evaluate the usefulness of this approach to significance testing.

These comparisons will be performed according to the following steps:

1. Determine whether a statistical test is needed for any individual item:
    - If preference for one of the devices on an item is unanimous, the pattern of results is clearly interpretable, and a statistical comparison between devices should not be performed for that item.
    - If ≥50% of the sample provided a neutral response for an item, a statistical comparison between devices should not be performed for that item. If <50% of the sample provided a neutral response for an item, proceed with steps 2 to 4 below.
2. Respondents who provide a neutral response for each item should be dropped from analysis of that item.
3. For each item, patients' responses should be grouped into two categories: (1) preference for Device 1 and (2) preference for Device 2 (in the current study, these devices are the semaglutide injection device and the dulaglutide injection device). These two categories of responses can be used to calculate the proportion who prefer each device. These two proportions are dependent, as they always add to 1.
4. Among patients indicating a preference for one device over the other, a statistical test can be performed to determine whether a significant preference exists:
    - This test would assess whether the proportion indicating preference for one of the two devices differs from 0.5. A value of 0.5 would indicate that there is equal preference for the two devices.
    - The frequency of preferences for Device 1 and Device 2 will be examined using a two-sided binomial test for each item of the DID-PQ.
    - For each DID-PQ item, the null hypothesis is that the probability of preferring one of the devices is 0.5. If the binomial test yields a significant p-value, then the null hypothesis can be rejected, which would mean that significantly more respondents preferred one device over the other.

These results will be presented in a series of three tables. The first will present results for the total sample. Then, the same significance testing will be performed within two subsets of patients: those who preferred semaglutide on the Global Preference Item and those who preferred dulaglutide on the Global Preference Item. It is expected that within each of these two preference subgroups, results of binomial tests will favor the preferred device. For example, patients who express a preference for dulaglutide on the Global Preference Item are expected to

be more likely to have statistically significant preferences for the dulaglutide device on items of the DID-PQ.

C. Aggregate Scoring of the Seven Device Characteristics Items of the DID-PQ

For the Device Characteristics items of the DID-PQ (i.e., items 1-7), another possible analysis approach is to perform a statistical test to assess whether one of the devices tended to be preferred more frequently than the other across the seven items.

This comparison will be performed according to the following steps:

1. For each individual respondent, compute three sums:
    - The number of items on which the respondent preferred Device 1
    - The number of items on which the respondent preferred Device 2
    - The number of items to which the respondent provided a neutral response (i.e., no preference between Device 1 and Device 2)

If all items were answered, the sum of these three values will equal 7.

2. Items receiving a neutral response (i.e., no preference between Device 1 and Device 2) should be excluded from the analysis.
3. For each respondent, compute a proportion of the frequency of items on which Device 1 was preferred divided by the total number of items on which a preference was indicated for either device. For example, if a respondent preferred Device 1 on four items and Device 2 on two items, with no preference for one item, then the proportion would be 0.67 (i.e., four preferences for Device 1 divided by 6 items on which either device was preferred). Patients who have all neutral responses will not be included in this analysis.
4. A mean of this proportion may be calculated for a patient sample. Then, a single sample t-test may be performed to assess whether this mean is significantly different from 0.5.
5. The null hypothesis would be that the mean proportion is 0.5. If the t-test yields a significant p-value, then the null hypothesis can be rejected, which would mean that significantly more respondents preferred one device across the Device Characteristics items.

These results will be presented in a series of three tables. The first will present results for the total sample. Then, the same significance testing will be performed within two subsets of patients: those who preferred semaglutide on the Global Preference Item and those who preferred dulaglutide on the Global Preference Item. It is expected that within each of these two preference subgroups, results of this aggregate scoring approach will favor the preferred device. For example, patients who express a preference for dulaglutide on the Global Preference Item should be more likely to have statistically significant difference favoring the dulaglutide device with this aggregate scoring approach for the DID-PQ.

The aggregate analysis described above can be conducted if <50% (i.e., three or fewer items) of the seven Device Characteristics items are missing. In this case, items that are not answered

would be excluded from calculation of the proportion. If >50% of the items are missing for any respondent, this respondent's data should not be used in the aggregate scoring approach.

6.12.4. Additional Exploratory Analyses

Additional exploratory analyses may be performed to further evaluate the individual factors associated with device preference. Logistic regressions with device preference as the dependent variable (i.e., sema vs. dula; or this could be dula vs. other response including preference for sema and no preference, depending on N in each preference group*)* may be performed to evaluate the impact of each DID-PQ item on device preference. If performed, model fit will be evaluated by receiver operating characteristic (ROC) curves as well as the Hosmer-Lemeshow goodness of fit test.

6.12.5. Other Analyses

6.12.5.1. Pilot Phase Analyses

Analyses of the pilot phase data will include both descriptive statistics and a content analysis approach. The purpose of the pilot phase is to ensure the training materials and interview procedures are clear, comprehensible, and feasible prior to the main phase.

Descriptive statistics (e.g., mean, standard deviation, frequency) will be used to summarize the sample in terms of sociodemographic characteristics and on the other measures completed by the participants (i.e., Global Preference Item, DID-PQ, four items selected from the MDDAB for each device).

A content analysis approach will be used to analyze the responses collected on the pilot phase interview questions. This content analysis applies only to the pilot phase. Responses will be examined to determine answers to the following questions for each device:

- Are there any ways the device training could be easier or more clear? If yes, please explain.
- Would you recommend any changes to the device training procedures? If yes, please explain.
- How much additional guidance is appropriate and useful during the device training?

During the pilot phase, the PI and study team will be continuously monitoring the interview and data as they are collected to determine if changes need to be implemented immediately. If major changes are required, Evidera and Lilly will determine if the pilot phase sample size needs to be increased. Upon conclusion of the pilot phase, Evidera study staff will send to Lilly a summary document of the pilot phase, which will serve as the interim analysis.

The pilot phase data for an estimated 10–20 participants will not be included in the final sample of approximately 290 patients with Type 2 Diabetes Mellitus.

6.13. Health Outcomes/Quality-of-Life Analyses

Not applicable.

6.14. Bioanalytical and Pharmacokinetic/Pharmacodynamic Methods

Not applicable.

6.15. Pharmacogenomic Methods

Not applicable.

6.16. Safety Analyses

This study focuses on preference between devices, and no medication will be administered to study participants. However, if a participant reports an adverse event (AE) related to the study, the appropriate reporting procedures will be followed. See Section 9.2 of the protocol for more AE details.

6.17. Subgroup Analyses

Not applicable.

6.18. Interim Analyses and Data Monitoring

No interim analyses will be reported but an interim dataset will be provided to the Evidera programming staff prior to data lock to allow programming work to begin before end of study.

6.19. Study Unblinding Plan

Not applicable.

7. References

Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423-429.

Gwet KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Fourth ed. Gaithersburg, MD: Advanced Analytics, LLC; 2014.

Matfin G, Van Brunt K, Zimmermann AG, Threlkeld R, Ignaut DA. Safe and effective use of the once weekly dulaglutide single-dose pen in injection-naïve patients with type 2 diabetes. *J Diabetes Sci Technol*. 2015;9(5):1071-1079.

Matza LS, Boye KS, Stewart KD, Paczkowski R, Jordan J, Murray L. Patient perceptions of non-insulin injection devices: Qualitative research to support development of the Diabetes Injection Device Experience Questionnaire (DID-EQ) and Diabetes Injection Device Preference Questionnaire (DID-PQ). *J Patient-Rep Outcomes.* In press; 2018a.

Matza LS, Stewart KD, Paczkowski R, Coyne KS, Currie B, Boye KS. Psychometric evaluation of the Diabetes Injection Device Experience Questionnaire (DID-EQ) and Diabetes Injection Device Preference Questionnaire (DID-PQ). *J Patient-Rep Outcomes.* In press; 2018b.

Matza LS, Boye KS, Currie BM, et al. Patient perceptions of injection devices used with dulaglutide and liraglutide for treatment of type 2 diabetes. *Curr Med Res Opin*. 2018;34(8):1457-1464.

Pictor A. Analysing Binary Outcome Data from a Crossover Design Study using the SAS® System. SAS VIEWS 2003; 8 pgs. Available at: https://www.lexjansen.com/views/2003/statpharm/st02.pdf.

Poon JL, Boye KS, Thieu VT, Norrbacka KN, Hassan SW, Gelhorn HL. Preferences for attributes of medications among patients with type 2 diabetes: a cross-medication class comparison of injection therapies. *Curr Res Diabetes Obes J*. 2018;6(5):555700.

Prescott RJ. The comparison of success rates in cross-over trial sin the presence of an order effect. *Appl Statist*. 1981 Jan;30(1):9-15.

Yu M, Van Brunt K, Milicevic Z, Varnado O, Boye KS. Patient-reported outcome results of dulaglutide added to titrated insulin glargine in patients with type 2 diabetes (AWARD-9). *Clin Ther*. 2017;39:2284-2295.

Leo Document ID = 6ca3362a-f941-4920-886e-b8a3d5715ed3

Approver: PPD
Approval Date & Time: 09-Oct-2018 02:03:49 GMT
Signature meaning: Approved